

Overheads cost research dear

Revelations about research universities that have charged the US government for yachts and other baubles have made new rules on indirect research costs inevitable. None too soon. But will the new rules be fair and flexible?

RESEARCH institutions in the United States receive some \$6,000 million a year from the National Institutes of Health (NIH) for the direct costs of biomedical research. The same institutions garner another \$2,000 million in overhead, euphemistically called indirect costs. Thus, NIH pay their share of the costs of running the university library and the president's office when they award a research grant. They pay for heat and light. Recently, the US Congress and public learned to their dismay that occasionally NIH (which is to say the taxpayer) also pay part of the costs of university yachts, antiques, fancy receptions and university-owned shopping malls. The revelations have been widely called a scandal. They are.

Nobody suggests that research institutions should (or could) meet out of their own resources the costs of providing an environment in which research projects can flourish. And it is unlikely that Stanford University, one of the most distinguished of research universities anywhere, is alone in having abused the system. But the university has given great offence by declaring at a congressional hearing that many of the challenged indirect cost charges were, by arcane accounting agreements, perfectly legal.

Now, government auditors have been despatched to review the books at a couple of dozen leading research universities, while the universities themselves have called in their own auditors. Harvard has announced already that it will withdraw at least \$500,000 in overhead charges; Massachusetts Institute of Technology (MIT) is withdrawing more than \$730,000 from its claims for indirect costs. According to MIT records, \$68,000 charged against overhead was spent on a Washington, DC, law firm that lobbied on the institution's behalf when Congress held a hearing on a paper co-authored by David Baltimore, then at MIT.

Outrage

Stanford has come in for the brunt of the criticism in this affair. The university's claim that some of its challenged charges were legal is an outrage to common sense. Television shots of its \$72,000 yacht and the president's elegant mansion have given an impression of academics as high-rollers, not tweedy or white-coated scholars. Stand-up comedians are having a field day at the university's expense.

But the scandal cuts much deeper. It has seriously damaged the image of universities nationwide as honest trustworthy brokers of public money. Furthermore, it reveals a system of government auditing in both the Department

of Health and Human Services and the Department of Defense — the two agencies that audit the books of research universities — that certainly needs repair. It is not that universities have conspired *en masse* to bilk the government. Rather, a system has grown up that allows them to recover some costs that auditors have approved but that, when brought to light, cannot possibly be justified as a cost of doing scientific research.

Tensions

What universities have lost in public trust is incalculable. What they stand to lose in dollars is not, and seems to be quickly adding up to millions. And this will hurt. It is important to remember that most indirect cost charges cover an unglamorous part of research — things such as heat, light, electricity, security and the library. The rules for allocating them are complex and make for reading that only an accountant could love. It has been hard to get anyone really to pay attention even though the issue has been simmering for some time, largely because indirect costs have been a rising proportion of direct research costs. Faculty at many universities have complained that they are being ripped-off by over-staffed and overpaid administrative offices, thereby creating unhealthy tensions between the faculty and university administrators.

Recognition that the system is in trouble has been widespread. The Association of American Universities produced a major analysis and set of recommendations only a couple of years ago. For the past 18 months, the association has been negotiating with the government's Office of Management and Budget to help reform the system, but the negotiations were suspended when Congress and the Bush Administration got into the act during the past few weeks. These manoeuvres nevertheless point the way to a likely resolution, one that will include a percentage cap across the board and new rules about which charges are allowable and which are not.

This is all to the good and long overdue. The scandal about yachts and antiques and lawyers' bills has had the salutary effect of focusing minds on an issue that deserves the attention it is now getting.

But a potentially more serious issue is the way in which the high costs of new buildings can be redistributed as indirect costs. The obvious danger is that zealous attempts to outlaw conspicuous abuse will generate inflexible rules that damage the research enterprise. □



Less class distinction

Plans to abolish a class difference in UK higher education is welcome; the need now is for autonomy.

THE British government has bitten a bullet any of its predecessors might have seized in the past quarter of a century, and has decided to abolish the distinction between universities and polytechnics (see page 257). The decision would have made sense at any stage. Its timing, now, is triggered by considerations as different as the demonstrated ineptness of the Universities Funding Council, whose scheme for parcelling out funds in response to commercial bids blew up in its face a year ago, and the wish to make education a comforting issue in the general election there must be in the next twelve months.

Whatever the explanation, the decision is correct. Twenty-six polytechnics were created (from pre-existing technical colleges) in 1966, at the instance of the late Mr Anthony Crossland and in the belief that they are cheaper than universities and in some ways better (in the diversity of their courses, for example). Over the years, they have evolved into an invaluable alternative source of higher education in Britain. Their formal dependence on local authorities was ended three years ago (the central government always wrote the cheques). Several of them have since acquired substantial reputations in research. Not merely is it natural that they should now be free to call themselves universities if they choose, but the resources available for higher education will be augmented, at no cost, if the disappearance of the snobbish concept of the binary system should tempt more young people to use the resources the polytechnics have to offer.

That is the good news. The bad news is that the same development will increase the competition for scarce research funds. Although the five British research councils are formally responsible for supporting research at both kinds of institutions, most of what they spend goes to, or on behalf of, traditional universities. The position is now further complicated by the planned transfer, next September, of funds intended to provide the indirect costs (see above) of academic research in universities. The plain truth is that the money now being spent on research and its overhead cost is not enough to meet the legitimate needs of the new claimants. But unless the claimants and the funds materialize, neither will the extra higher skill that the government rightly asks for. M. Lionel Jospin in France has a similar problem (see below).

That is why the success of the merger now planned will hang on the intelligence with which higher education is supported financially. There are to be separate sources of funds for higher education in England, Scotland and Wales, but little else has so far been said. The present government is now being attacked for the autonomy it has given a number of National Health Service hospitals, ostensibly so that they should be free to manage their affairs efficiently. It is a long time since British universities, which are titularly self-governing institutions, were that free. If the intended benefits of

the merger are to materialize, it is essential that the institutions concerned should be enabled to make their own way in the government's new world — and even run the risk of failing from time to time. Is that what is intended? □

Research assured

The new prime minister of France is likely to be even more supportive of research than her predecessors.

M. FRANÇOIS Mitterrand is running through prime ministers as if they were paper tissues. After six months of trench warfare in the National Assembly, M. Michel Rocard has been let go, to be replaced by Mme. Edith Cresson, whom *Le Monde* has called the 'velvet lady' in mocking emulation of Mrs Margaret Thatcher's common nickname. But by all accounts (including that of *Le Monde*), the velvet may be only glove-deep. That is just as well, for the trench warfare will not go away, and may intensify as legislative elections approach. The underlying difficulty is not so much electoral as that the French Socialist Party is being increasingly factionalized by the need to adapt to changing circumstances.

What this change will mean for the condition of science in France remains to be seen, although the signs are encouraging. The two ministers most directly concerned, M. Hubert Curien at the ministry of research and M. Lionel Jospin at the ministry of national education, remain in their places. For almost a decade, Curien's influence on French research policy has been decisive, consisting as it has done not merely in the steady growth of the funds available, but in a steady reinforcement of people's enthusiasm. Jospin necessarily skates on thinner ice. National policy aims at putting two out of five young French people through some form of higher education, in emulation of Japan among other industrial states. The difficulties are not only physical (building universities quickly enough), but qualitative: as in comparable societies, Britain for example, knowing how and even where to strike a balance between the general spread of modest skill and the narrower cultivation of greater expertise is never simple. Jospin has his work cut out.

Mme. Cresson should help, to judge from the questions she has been raising in her first few days. Why does France train fewer engineers than Germany for example? But the plan to make M. Pierre Bérégovoy, who remains as the minister of state with responsibility for the economy, finance and budget, into a coordinator of industrial policy as well is a wilder card. At the back (even the front) of many minds is an ambition to create an alliance between government and industry as influential as Japan's Ministry of International Trade and Industry was supposed to be in the 1960s. One danger is that of backing losers, not winners. Another is the risk of entrenching French chauvinism when much of Europe is moving in the other direction. The troublesome left-wing of the Socialist Party may be mollified, but what will matter in the long run is whether the recipe satisfies France's unique blend of radicalism and conservatism. □

University/polytechnic distinction to end

- Little change seen in research funding
- Education on the election agenda

London

THE division in British higher education between the universities and polytechnics should soon end. Saying that "The polytechnics have come of age," Prime Minister John Major launched on Monday (20 May) a series of reforms to education and training, including a plan to eliminate the distinction between the two types of institutions.

The first polytechnics were set up in the 1960s to provide vocational education aimed at preparing students for jobs in industry and other professions. But now they are often seen as 'second class universities' by prospective students and employers alike.

Under Major's new proposal, the Universities Funding Council (UFC) and the Polytechnics and Colleges Funding Council (PCFC), both created in 1989, would be dissolved. They will be replaced by three new higher education funding councils, one each for England, Scotland and Wales, which will find both teaching and research across the whole higher education sector. Polytechnics will be able to call themselves universities, should they wish to do so.

The abolition of the traditional 'binary divide' in British higher education will allow the polytechnics to compete for students on level terms with the universities, Major said. And Sir Ron Dearing, PCFC chairman, and Lord Chilver, his counterpart at the UFC, issued a joint statement on Monday welcoming the proposed changes.

The decision by Major, who left school at sixteen, to launch the new initiative personally suggests that education will be a central issue in the next general election campaign. The opposition Labour party has already stated that the UFC and the PCFC would be merged, if Labour comes to power.

Many polytechnic directors have been pushing hard for the merger of the universities' and polytechnics' separate funding bodies (see *Nature* 349, 360; 31 January 1991). The directors would like some of the £860 million the UFC distributes each year for research in the universities to be made available to polytechnics with a good record for research. (The PCFC's research budget in 1990-91 was only about £30 million.)

The proposed new system, however, is unlikely to result in a major redistribution of British research spending. Education and Science Secretary Kenneth Clarke said there would be no wholesale transfer of research funds from the universities to the polytechnics. Under the system, the former polytechnics will merely be able to compete "in a limited way" for research money from the new funding councils, he said.

The higher education policy document does not rule out more radical reform of British research spending in future. In the long term, the research budgets of the new higher education funding councils may be transferred *en masse*, either to the research councils, or a new research funding agency.

Many polytechnics are expected to relaunch themselves as universities, once they are able to do so. But the change in title may be little more than a cosmetic exercise. Clarke said the new funding councils will be asked to ensure that the polytechnics do not change their mission. The former polytechnics should still concentrate on vocational education and applied research, the universities on academic subjects and pure research, he said.

The new system will also do little to silence the critics within the universities and polytechnics who claim that the government's policy of expanding higher education without a commensurate increase in public funding will erode the quality of British higher education. But the higher education policy document does propose that a new independent unit should be set up to monitor the quality of education across the higher education sector.

The package of policy documents released on Monday also include reforms to the education and training of sixteen-to nineteen-year-olds. The aim, according to Major, is to "break down the artificial barrier which has for too long divided an academic education from a vocational one". Young people will be free to take a mixture of A-levels, the traditional academic qualifications which form the standard entry requirement for higher education, and vocational qualifications.

But the government has not bowed to the calls for the replacement of A-levels with a system that does not channel students into a narrow range of academic subjects at sixteen. (Students in England and Wales aiming to enter higher education typically take three or four A-levels, each in a single subject.) Earlier this month, the Royal Society suggested that A-levels should be replaced by a single qualification within which students would be able to take a range of course modules, from a diverse range of subjects but the government rejected this.

Major said that the government would introduce legislation to enact the proposed reforms as soon as possible. But with a general election possible before the end of the year, the government may have to be re-elected before the new policies are put into action.

Peter Aldhous

SPACE SCIENCE

Astro reborn

Washington

AFTER a lobbying effort by Senator Barbara Mikulski (Democrat, Maryland), the National Aeronautics and Space Administration (NASA) has agreed to fly the "Astro" package of three ultraviolet telescopes on at least one more shuttle mission.

The package's first outing, called Astro 1, was aboard the space shuttle Columbia last December. Despite a series of mechanical troubles and mishaps, it was operated almost continuously for 9 days by shuttle astronauts, who were guided by instructions from astronomers at Marshall Space Flight Center in Alabama (see *Nature* 348, 571; 13 December 1990). The three telescopes obtained images and spectral and polarization data in ultraviolet wavelength bands inaccessible to ground-based astronomers; they were accompanied by a separate X-ray telescope which will not be flown again on Astro-2. The first scientific results from the Astro-1 mission were published in *Nature* two weeks ago (351, 128; 9 May 1990).

In its original conception, the Astro package was designed to be flown at least six times, but after the inaugural launch was delayed first by the Challenger disaster of 1986 and then, in early 1990, by a string of shuttle problems, it was unclear whether the demands on the shrinking shuttle launch manifest would permit it ever to be flown again. Mikulski, whose interest in Astro derives from the fact that two of its three telescopes were built in Maryland (at Johns Hopkins University and the Goddard Space Flight Center), intervened to persuade NASA to take Astro out of storage and put it back on the shuttle launch manifest.

David Lindley

FRENCH RESEARCH

Science policy from abroad

Washington

EDITH Cresson, the new prime minister of France, wants her country to have as much government support of industrial science as Japan and as many engineers as Germany. In one of her first acts in office, last week she consolidated the French ministries of finance, economics, foreign trade, industry and telecommunications into one "super-ministry" in a bid to emulate the power and influence exercised by the Japanese Ministry of International Trade and Industry.

She reappointed Pierre Berégovoy, finance minister, to lead the new agency, which will be the focus of the new prime minister's ambitious industrial policy — to fashion France into an economic power on the level of Germany.

Christopher Anderson

FDA upgrade proposed

Washington

THE US Food and Drug Administration (FDA) is so "overextended, underfunded, and shackled by bureaucratic restraints" that it should be taken out of the huge Department of Health and Human Services (HHS) in which it is submerged and given new status as an independent agency, perhaps along the lines of the National Science Foundation. This is the chief recommendation of an *ad hoc* FDA review committee that was appointed a year ago by HHS Secretary Louis Sullivan and which reported back to him last week.

Sullivan was not pleased. The committee, headed by former FDA commissioner Charles C. Edwards, now head of the

tion to free the FDA if the Administration refuses to budge.

A strong FDA is vital to many segments of the US scientific enterprise — the biotechnology industry in particular. During the past four years, the length of time it takes overworked FDA staff to complete the review and approval process for new drugs has increased from two years to 34 months. With the exception of AIDS-related therapies, the increase in the number of new drugs awaiting final review has not been accompanied by a material increase in the number of scientists in the FDA's Center for Biologics Evaluation and Review, which handles 80 per cent of all biotechnology products. This, the committee argues, could have serious implications for the international competitiveness of US industry. The analogous review process

in the European Committee for Proprietary Medicinal Products takes about a year.

The committee urged the FDA to adopt new and more flexible standards for approving drugs, particularly those for life-threatening diseases and diseases for which no therapeutic alternative to a new drug exists. This, it believes, could be best accomplished if the FDA commissioner had real decision-making authority and could act more independently.

Although the Edwards group sees independence and authority as the key issue for FDA, more money is needed too. In the current climate, with large increases of federal funds unlikely, the idea of charging companies 'user fees' to deal with FDA is being revived. In fact, the Administration's proposed budget of \$770 million for fiscal year 1992 assumes that users will pay fees of \$197 million — money that could be used to recruit new scientific staff. **Diane Gershon**

SUPERCONDUCTING SUPER COLLIDER

New detector sought

Washington

IT'S back to the drawing board — literally — for a particle detector to be used on the Superconducting Super Collider (SSC).

Last week, Roy Schwitters, director of the SSC, announced that he will not pursue the so-called L* (pronounced 'L-star') detector and invited physicists from around the world to help with the development of a detector to take its place.

The consortium proposing the L* contained groups from 13 countries' and its downfall threatens to spoil the international flavour of the SSC. But Schwitters says the lessons learned from the L* fiasco should make it easier to foster a new international collaboration.

Two major detectors, each costing \$500 million or more, are planned for the SSC, whose projected cost is itself \$8,250 million and rising. A group headed by George Trilling of Lawrence Berkeley Laboratory in Berkeley, California, received the go-ahead in January to develop a full design proposal for the primary detector. At the same time, the SSC Laboratory narrowed its choices for a second detector to one proposal, which had been submitted by a consortium of 90 institutions headed by Samuel Ting of the Massachusetts Institute of Technology. Ting's group was invited to submit an amended proposal; in particular, the SSC Laboratory wanted Ting to allay its concerns about L*'s cost, its management structure and the strength of the US constituents of the consortium.

An independent cost estimate completed in March concluded that L* would cost more than the consortium estimated, but not enough in excess to cancel the proposal. Ting, however, never adequately addressed the questions about L*'s management, Schwitters says, and after Germany, Switzer-

land and the Soviet Union withdrew from the consortium in March and April, the loss of their combined \$200 million commitment made the cost of the L* an issue once again.

The four-month debate over L* has left some hurt feelings, especially in Europe. Particularly galling to many was the SSC Laboratory's conclusion that L* needed more and stronger US participation. Schwitters says this reflected concern over the number of US universities in the consortium that were not traditionally strong in high-energy physics and was not meant to denigrate the European members of the consortium. But some of the European collaborators took it this way, and Schwitters now admits that the wording was 'insensitive'.

From 11 to 13 June, the SSC Laboratory will host an open workshop to discuss the development of another detector, and Schwitters says he will work to regain the trust of the international physics community which has been damaged by the laboratory's handling of L*.

Switzerland, home of CERN, the European particle accelerator, may not participate this time around, Schwitters says — Swiss officials were angry with the handling of the L* consortium, and CERN would prefer that Swiss institutes spend their money inside the country. But Schwitters has high hopes for Germany and the Soviet Union, insisting that the L* experience has 'solidified' ties with Soviet scientists in particular.

"Assuming some interest develops at the meeting, we'll pick up the pieces" and go on, Schwitters says. With luck, the whole experience will set back the timetable for detector development less than a year, and both detectors will be ready when the SSC comes on line, now scheduled for 1999.

Robert Pool

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

FDA advisory committee seeks more clout for new FDA commissioner, David A. Kessler.

Scripps Clinic and Research Foundation in California, was blunt in saying that FDA should be freed from HHS. Giving the commissioner real authority to issue regulations, recruit senior staff and manage the FDA's internal affairs without interference from HHS officials who have other things on their minds "would do more than any other single measure available to HHS to restore the commissioner's prestige, enhance the agency's effectiveness, and improve employee morale," the 15-member committee concluded.

Midway through the committee's deliberations, David A. Kessler stepped down as a committee member when he was appointed FDA commissioner by Sullivan.

Sullivan countered the committee's recommendations, predictably, by saying that it would be a serious mistake to separate FDA from his department in view of the need for close coordination between FDA and other HHS agencies, particularly the Centers for Disease Control and the National Institutes of Health.

However, Senator Edward M. Kennedy (Democrat, Massachusetts) is siding with the Edwards committee on this. He said that the committee put forward a "compelling case" and that he is prepared to introduce legisla-

Whaling ban versus science

London & Tokyo

NEXT week, an annual ritual will be acted out in Reykjavik, Iceland. At a meeting of the International Whaling Commission (IWC), the language of science will be used to debate a question that actually is dominated by ethical and economic concerns: should the great whales be hunted commercially?

On one side stand the anti-whaling IWC members, whose populations generally disapprove of the killing of whales — a majority that includes the United States and Britain. On the other stand Iceland, Norway and Japan, where coastal communities that once relied upon whaling for their livelihoods have been forced into economic decline by the present moratorium on commercial whaling.

The IWC, however, was not set up to address the ethics of whaling, nor to ensure the economic well-being of whaling communities, but rather to regulate whaling according to scientific principles so as to conserve whale populations. In the past, anti-whaling nations have used uncertainties in the understanding of whale population biology to block any commercial whaling, but IWC scientists are now on the verge of producing a computer model designed to set catch quotas that do not endanger whale populations. This accomplishment will make it extremely difficult to oppose all commercial whaling within the framework of the IWC.

The current commercial moratorium on whaling came into force in 1986 (although Norway and Japan continued commercial whaling for another two years). The argument behind the ban was that the techniques the IWC had been using to set quotas for whale catches were inadequate: there was a danger that whale stocks could be seriously depleted, even by catches within the limits set by the IWC.

Since the ban has been in force, the IWC's scientific committee has been reassessing the size of whale stocks around the world, and working on a 'revised management procedure' to replace the old quota-setting system. The whaling nations are impatient for this to be put in place, so that they can resume commercial hunting of the minke whale, the smallest but most numerous of the great whales.

Biologists on the scientific committee have been developing five separate computer models based on whale population dynamics. They have been asked by the IWC to produce a single model that can be used to set quotas that are large enough and stable enough from year to year to support a commercial whale fishery, yet that will still protect whale stocks from depletion. The biologists hope to decide on a single model in time for next week's IWC meeting.

Once a revised management procedure is in place, the argument in favour of maintaining the ban will be difficult to sustain under

the IWC's constitution. The IWC is for "the regulation of whaling, not the prohibition of whaling," says Jan Arvesen, Norway's IWC commissioner.

The scientific committee is now meeting in closed session in Reykjavik. The full IWC, meeting from 27 to 31 May, must decide whether to adopt the management procedure put forward by its scientists. If it does so, the scientific committee must then conduct further detailed simulations with the model before it can be used to set real catch limits. This could take at least another year.

US and British delegates to the IWC meet-

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Can Minke whales be hunted commercially without endangering their populations?

ing are under public pressure to block the adoption of a management procedure, which would open the door to future commercial whaling. Some whaling nations are worried that this pressure may lead delegates from the anti-whaling countries to throw up various objections to the scientists' computer models in order to delay the adoption or use of any new management procedure.

Kevin Chu, from the US State Department, denies that the US delegation intends to "hide behind the science". The IWC scientific committee has made good progress towards developing a workable computer model, he admits, but he argues that more work must be done before the model is used to set quotas for real catches.

The whaling nations also expect to hear questions about whether the methods used to kill whales are humane — a subject that raises the hackles of the Icelanders, Norwegians and Japanese. "Are killing methods in the slaughterhouses of Chicago humane?" Arvesen retorts. "If they think there is a humane method of killing cows, we will take up the same method for whaling," says Fukuzo Nagasaki, scientific adviser to the Japanese delegation.

The United States may also find its role as the 'world policeman' of the whaling issue undermined by events in Reykjavik. Alone among the anti-whaling nations, the United States has legislation that can be used to apply economic sanctions against countries defying the IWC, or endangering whale populations. But the US delegation must this

year reapply to the IWC for permission for its Alaskan Inuit population to catch 40 bowhead whales a year for the next three years. This 'aboriginal subsistence whaling' is not affected by the commercial moratorium, but catches must be approved by the IWC. As the bowhead is an endangered species, US officials are bracing themselves for some opposition to their request.

Chu defends the US bowhead catch, saying that the stock is one of the best studied whale populations, and has been growing at a rate of three per cent a year, despite the annual Inuit catch.

Chu's argument does not necessarily contradict his assertion that the knowledge of minke populations is insufficient to sanction

a resumption of commercial whaling — the positions of the boundaries between different minke stocks are not well defined. But the apparent inconsistency will make an appealing target for the whaling nations, particularly for Iceland, the nation now pressing most strongly for a resumption of whaling. Iceland intends to ask for an interim commercial quota for 192 minke whales (from a stock of some 28,000 whales) and 91 fin whales (from a stock of 15–16,000)

from the central North Atlantic over the coming year. Judging from the response to previous similar requests, this is unlikely to be accepted.

The debate may not come to a head this year. The anti-whaling nations want to delay any decision on the management procedure, and the scientific committee may not be able to choose a computer model this year. But the desire to defer the issue must be weighed against the longstanding threat by the whaling nations to leave the IWC.

Arni Kolbiensson, secretary-general of the Icelandic fisheries ministry, says that Iceland joined the IWC as a "resource utilization organization". If the IWC fails to conform to that description, he says, Iceland will be forced to reconsider its membership. Whether this threat proves to be more serious than those made in previous years remains to be seen.

Nagasaki is convinced that some IWC nations are using the development of the management procedure as a delaying tactic. And some members of the scientific committee wish privately that the IWC's constitution would allow the true argument — the ethical concerns of the antiwhaling countries versus the economic agenda of the whaling nations — to be debated openly. As one committee member observes, it would be "about time, and much more honest", if the delegates next week in Reykjavik focused on the real issues that concern them, rather than the finer points of whale population dynamics.

Peter Aldhous & David Swinbanks

Tokyo gears up for calamity

Tokyo

THE Tokyo metropolitan government is spending hundreds of millions of dollars on the latest technology designed to manage disaster relief operations if Tokyo is hit by a major earthquake or typhoon.

The metropolitan government moved last month from old buildings in the centre of Tokyo to a new 48-storey skyscraper in the northwest of the city. The building — Tokyo's tallest — cost ¥144,000 million (more than \$1,000 million) and bristles with high-technology surveillance equipment that will gather information and direct rescue activities in the event of a disaster.

Tokyo lies on the edge of three colliding plates — the Pacific, Eurasian and Philippine Sea plates — and a major earthquake could hit the city at any time. In 1923, the Great Kanto Earthquake struck Tokyo, killing 140,000 people, and although a repeat is not thought to be imminent, several experts do think that such an earthquake could strike in the first few decades of the next century. Tokyo also lies in the path of typhoons that batter Japan in autumn and sometimes cause serious flooding in low-lying areas of the city.

Thus disaster prevention is serious business for Tokyo, and the equipment in the new building is correspondingly state-of-the-art. On the roof are two panoramic television cameras on the east and west sides that can each scan through 180 degrees, giving a complete 360-degree coverage of the city. The cameras have lenses that on a clear day allow government officials to zoom in on buildings in Tachikawa city, 30 km away.

The aim is to pinpoint fires that will break out all over the city if a major earthquake strikes. In the 1923 earthquake, it was fire storms that caused most of the 140,000 deaths.

In addition to those on the roof, special

television cameras mounted on two helicopters operated by the Tokyo Fire Department and the Tokyo Metropolitan Police Department will be sent into the air to scan the city when disaster strikes. The cameras, mounted on arms off the side of the helicopters, are designed to remain level no matter how the helicopter rocks and they automatically orientate themselves so that north will



Awaiting disaster

always be towards the top of the television picture. They can also take infrared pictures at night.

At that point, the most important (and expensive) part of the disaster prevention system comes into play. The top half dozen floors of the city headquarters bristle with microwave radio antennas pointed at 23 local ward offices around the city as well as at national government ministries, police and

fire departments, headquarters of the self-defence forces, television stations, telephone, electricity, gas and water utility companies. All in all, more than a hundred organizations in Tokyo are linked to the microwave radio communication system which can operate when ordinary telephone lines are broken by an earthquake or flooding.

This communication system, which cost about ¥20,000 million (\$150 million), will be used to direct relief and evacuation operations and to gather the latest data on the state of the disaster from around the city.

Operations will be directed by the governor of Tokyo and other officials who will man the disaster prevention headquarters located between the eighth and ninth floors. Three huge screens on the wall of the headquarters will display live pictures from the television cameras and maps of the city showing the latest statistics on the spread of fires (or floods) and deaths in the city. Orders will be issued through a nearby communications room equipped with telephones and facsimile machines connected to the microwave communication system.

All of this preparation, however, will do little for the people who live and work in the areas around Tokyo. And if a major earthquake or flooding caused by a typhoon hits that city, it is almost certain to affect areas beyond the city limits. For example, the area of the Izu Peninsula, 100 km south of Tokyo, is one likely location for the epicentre of a major earthquake to hit Tokyo.

In such circumstances, coordination of disaster relief operations becomes the responsibility of the national government, which has not coordinated its disaster response as well as the city of Tokyo. Almost every national government ministry or agency has a section dealing with natural disasters, but there is no central organization in charge of them all. And expert observers express scepticism about the ability of all the various government ministries and agencies to act quickly and effectively in the event of a disaster.

For example, Peter Hadfield, a geologist, points out that when a Japan Airlines jumbo jet crashed into a mountain north of Tokyo early one evening in August 1985, it took hours for the Transport Ministry, which is in charge of the airline industry, to issue a request for help from the self-defence forces, which are under the jurisdiction of the Defense Agency. And no government response was given to offers of help from US Air Force bases in the outskirts of Tokyo that have helicopters equipped with night-viewing cameras. Rescue teams in helicopters were not despatched to the crash scene until early the next morning, by which time most of the survivors of the crash had died.

David Swinbanks

Worldwide costs of an earthquake

Tokyo

If a major earthquake strikes Tokyo, the Japanese will not be the only ones to feel the pain. According to a report by Tokai Bank in March 1989, large-scale damage to this world financial centre would throw other parts of the world — particularly the United States — into a severe recession.

The bank's report examines the consequences of a repeat of the Great Kanto Earthquake of 1923, which killed 140,000 people. In one scenario, the report predicts that such an earthquake could cause \$650,000 million in damage to the Tokyo area, equivalent to nearly a quarter of Japan's gross national product.

In such a case, the report predicts, the Japanese economy would first be plunged into a short period of negative

growth. But then Japanese investors, especially insurance companies, would quickly pull money out of investments overseas, which are concentrated largely in the United States, to pay off liabilities and to rebuild the Tokyo area. Japan would then experience a temporary economic boom during reconstruction of the city, just as it did in 1923, but the United States would be thrown into economic recession. (Japan provides nearly 20 per cent of foreign capital inflow to the United States, according to the Tokai Bank report.)

With the US recession would come a stock and bond market crash. And a consequent sudden rise in US interest rates would adversely affect US trading partners and developing nations that are heavily in debt to the United States, according to the report.

D.S.

German proposal draws fire

Munich

IN order to help Israel cope with the influx of thousands of Soviet Jewish researchers, Germany last month announced a new DM1 million (\$592,000) programme to enable some of those researchers to visit Germany for stays of up to six months. The purpose was to acquaint them both with German scientists and with the techniques of modern research.

But a number of Israeli researchers are criticizing the programme as the wrong way to assist Israel. "Why can't they subsidize Soviet Jews who will be staying in Israel?" asks Yoram Milner, a professor of biological chemistry at Hebrew University in Jerusalem. Milner says the German programme is ill-conceived and sounds "a bit bizarre".

Michael Schlesinger, a professor of medicine at Hebrew University, is more subdued in his comments, but still critical. "In general, this is a worthwhile project," he says, but it is "not the first thing I would have done" with the money. Israel urgently needs funds to absorb the researchers at home, he explains.

Underlying the discord seems to be the unspoken conviction that Jewish scientists who leave the Soviet Union should go to Israel and nowhere else, and German officials know that this is a touchy subject. The German Research and Technology Ministry,

HIGH-ENERGY PHYSICS

HERA stores protons

Munich

THE 820-GeV proton ring of the HERA electron-proton collider at DESY (Deutsches Elektronen-Synchrotron) was successfully run with its first protons last week. Several 'packets' of protons were stored in HERA at the injection energy of 40 GeV.

The next step will be to bring on line the superconducting high-frequency cavities used to raise the energy of the electrons in the ring from 27 GeV to the final energy level of 30 GeV. (The electrons circulate in the same ring as the protons, but in the opposite direction, so that they can be directed to collide with each other, with physicists observing the resulting debris of the collisions.) After that, the energy of the protons will be gradually increased in test runs until the final energy is achieved.

If all goes according to plan, the first collisions will come by November, when the two detectors ZEUS and H1 are due to come on line.

HERA (Hadron Electron Ring Anlage) took six and a half years to build and cost approximately DM1,000 million (\$592,000). It is the world's first high-energy electron-proton collider and the first European accelerator to use superconducting magnets. Eleven countries contributed to building HERA, which is located in the Bahrenfeld section of Hamburg.

Steven Dickman

for instance, says it is aware that the *émigrés* might attempt to stay in Germany or somewhere else in Europe, where the job market is bound to be better than it is in Israel. Therefore, says Ministry official Peter Gottstein, Germany intends to let Israel choose the researchers who are to come to Germany from among those who already have jobs in Israel. "We're not out to do headhunting here," says Gottstein.

Yael Shinar, a spokeswoman for the Israeli Ministry for Science and Technology, says that there might be changes in the "mode of operation" of the programme, which might allow Soviet *émigrés* to use some of the money without visiting Germany. But the German Ministry made no allowance for this when it announced the programme during a visit by German Research Minister Heinz Riesenhuber to Israel in April.

The Israeli science ministry estimates that of the 300,000 *émigrés* in the latest wave from the Soviet Union, some 2 per cent, or 6,000 people, are in scientific research (not

GERMANY

New university for Frankfurt on the Oder?

Munich

LED by Peter Wex, a legal scholar and administrator at the Free University of Berlin, a group of academics and politicians in Germany is intent on reviving a university in the eastern German city of Frankfurt on the Oder. That city had a university for more than three centuries until 1811, when it was closed during the Napoleonic wars.

The proposed university, Wex says, could take advantage of Frankfurt's proximity to Poland to offer courses in Eastern European studies and to provide a German-language education to Poles and other Eastern Europeans. Frankfurt, a border town with 87,000 inhabitants, has been in the news lately as the site of brutal attacks on Poles by German nationalist and neo-Nazi groups.

Wex's group seems to have won the sympathy of the influential science advisory council Wissenschaftsrat. Reestablishing a university in Frankfurt is "an idea that could not be more welcome", says spokesman Wilhelm Krull. Support from Wissenschaftsrat would be crucial in gaining support from Bonn if the idea for a university in Frankfurt is backed by the government of Brandenburg, the *Land* (state) in which Frankfurt is located. (The city is 500 km east of its namesake in the west, Frankfurt on the Main.)

But Brandenburg may take its time in backing Wex's proposal. Although Brandenburg's governor, Manfred Stolpe, announced in his inaugural speech in

counting engineers or medical doctors). They represent an almost insurmountable absorption problem for a country as small as Israel, which in mid-1989 had an estimated 4.37 million inhabitants.

Both the Israeli researchers and ministry officials interviewed are eager not to antagonize Germany, which over the years has spent DM400 million for cooperative projects with Israeli science. Scientific relations between the countries, which go back to 1959, even predate diplomatic relations, which began only in 1965.

Germany, too, is eager to maintain a positive image in a country where people are inherently suspicious of its motives. "We have to keep persuading [the Israelis] of our goodwill," Gottstein says.

Nevertheless, Milner says the Germans have made the wrong move in this case, especially in the aftermath of the Persian Gulf war. "Germany has more work to do than before on its image," he says, because of the role played by German companies in delivering supplies to Iraq before and during the Gulf War. "The Germans need to practise more caution and restraint."

Steven Dickman

January that he plans to found three universities in Brandenburg, including one in Frankfurt, he may not have the money. Krull says that "three or even two universities" would be a big burden for Brandenburg (which currently has none), as the *Land* would ultimately have to finance them from its own tax revenues.

Stolpe has indicated that Brandenburg's capital city, Potsdam, is the preferred site for a university, which would be pieced together from the remnants of a pedagogics institute and an institute for law and administration.

Nevertheless, Wex is canvassing for Frankfurt as the site of at least a small university. A new university would not be burdened by the ideological ballast of the past, he points out.

Also — and here Krull concurs — it would be less expensive to start up a university that possessed faculties for only law, economics and social sciences, leaving the more expensive natural sciences for later.

Finally, Wex adds, western companies have expressed an interest in endowing chairs at such a university. But speed is of the essence, he warns. "They won't wait around forever."

Oddly enough, says Wex, he has so far received little support from those who would benefit the most from a university in Frankfurt: the city's residents. He attributes this to the general struggle for existence that has overtaken the citizens of eastern Germany.

Steven Dickman

From hillside to laboratory

London

LESS than two weeks ago, botanist Riadh Dinha Francis and his family were living in the Shemdinly refugee camp on the Iraqi/Turkish border. This week, following a concerted effort by British scientists and diplomats, he is in Sheffield, ready to resume research interrupted by the war in the Persian Gulf.

Francis owes his rescue to a chance appearance on British national television. Interviewed for Channel 4's nightly news bulletin in April, he was recognized by Terry Croft, manager of the Department of Animal and Plant Sciences at the University of Sheffield. Francis had completed his doctorate at Sheffield in 1985, and then returned to Salahaddin University in Arbil, Iraq. He hoped to return to Britain, however, and before the outbreak of the war, he had applied for a Royal Society fellowship to continue his research at Sheffield.

After spotting Francis on the broadcast, Croft asked the Royal Society and Channel

4's science correspondent, Andrew Veitch, to help find the Iraqi researcher and bring him to Britain. Several weeks later, after intense behind-the-scenes effort by British diplomats, Francis and his family were flown from Turkey by the Royal Air Force.

Francis, an Iraqi Christian, says that most of Salahaddin University's staff and its 6,000–7,000 students fled from Arbil when the Iraqi army advanced against the Kurdish rebels: "Many of my students are on the mountains," he says.

Salahaddin, the only university in the Kurdish region of Iraq, has effectively been closed since the beginning of the Gulf War, Francis says. During the Kurdish uprising, he reports, the college of education was "completely destroyed" by Kurdish rebels venting their anger against the Baghdad regime. The rest of the university is intact, he says, but he doubts that many of its staff will ever return.

Conditions in the refugee camps on the Iraqi/Turkish border were beginning to improve by the time he left, Francis says,

although the distribution of food and medicines was still chaotic. When he was interviewed in April, Francis and his family had not eaten for three days.

Francis also corroborates stories that have emerged from Kuwait of the theft of scien-



Riadh Dinha Francis with Terry Croft, his rescuer.

UK FORENSIC SCIENCE

Forensic reform needed now

London

IN the wake of the celebrated release of the Birmingham Six, experts are saying that Britain needs to rethink the way it provides forensic science testimony to the courts.

In March, the conviction of the Birmingham Six was overturned after the forensic evidence used to convict the six Irishmen accused of two terrorist bombings in Birmingham in 1974 was discredited. That case has prompted a Royal Commission that will consider reforms to the British judicial

lawyers representing suspected criminals in British courts must often prepare their cases without adequate scientific advice. Unless the many "cowboys" operating as independent forensic scientists are regulated, Gallop said, the treatment of scientific evidence in British courts will remain biased in favour of the prosecution.

The problem with the present British system, Gallop said, is that the same Home Office scientists who have been intimately involved in a police investigation are then asked to stand in court and give dispassionate scientific evidence. Asking Home Office forensic scientists to "pick holes" in their own evidence, said Gallop, is like "expecting a hound which has just caught a hare . . . to give it the kiss of life".

Independent forensic scientists, Gallop said, are vital to help prepare defence cases, and to give evidence in court. In the absence of regulation, any person with a scientific background can set himself up as an independent forensic science consultant.

Gallop called on the Forensic Science Society and the Law Society to rid the profession of such unqualified practitioners.

Home Office and police laboratories now have a virtual monopoly in training Britain's forensic scientists, and too few scientists leave the Home Office to work independently, Gallop said. The reason, she said, is the long delays between cases coming to court and independent forensic scientists getting paid. The legal fees for many defendants are paid by the state, through the system of Legal Aid. But this often takes many months, or even years to process.

Peter Aldhous



system. But the Royal Commission is expected to take several years to gather its evidence, and at least one forensic expert says that reform of the forensic science profession is needed much sooner.

Angela Gallop from Forensic Access, one of only a handful of respected British independent forensic scientific consultants, told a meeting at the Royal Society this week that

tific equipment during the Iraqi occupation. "We have seen lorries filled with equipment from Kuwait University," he says. But he adds that the theft was organized from Baghdad — the vice-chancellor and chancellor of Salahaddin were not involved.

Peter Aldhous

SPACE

First Briton hits orbit

London

At 15:50 Moscow time on 18 May, Helen Sharman blasted off from the Baikonur launch site in the republic of Kazakhstan to become the first Briton in space. Two days and more than 30 orbits of the Earth later, the Soyuz capsule carrying Sharman and her two Soviet colleagues docked successfully with the Mir space station. Sharman will spend six days aboard Mir, conducting experiments devised by scientists working for the Soviet space programme, before returning to Earth with two Soviet cosmonauts, who have been working on board Mir for the past six months.

The Anglo-Soviet Juno mission, which aimed to put a British astronaut and a roster of British microgravity experiments into space, came near to collapse last year when commercial sponsors failed to come forward. The mission was saved when its original financial backers, the London-based Moscow Narodny Bank, offered to pay the Soviets an undisclosed sum to secure a flight for a British astronaut. The sum involved is thought to be much less than the mission's original estimated cost of £16 million. At a pre-launch press conference, Soviet journalists are reported to have expressed dissatisfaction at the Soviet Union 'subsidizing' Sharman's flight.

Peter Aldhous

IQ and vitamin supplements

SIR — None of Steven Blinkhorn's criticisms (*Nature* 350, 13; 1991) of the study of IQ improvement following vitamin and mineral supplementation¹, financed by the Dietary Research Foundation, are acceptable. His chief criticism is directed at the statistical treatment. He argues that significance tests should have been carried out on all the tests and supplementation formulae issued. But this idea ignores the fact that we tested specific hypotheses, rather than proceeding blindly. We hypothesized, on the basis of past findings and psychological theory, that tests of fluid ability (nonverbal tests) would show an effect, while tests of crystallized ability (verbal tests) would not. We made the Wechsler (WISC-R) individually administered test our main tool, as any psychologist would do; the data bore out both hypotheses. It would have been quite nonsensical to count the failure of the verbal tests against our hypothesis; this is exactly what was expected.

We also included some tests of scholastic achievement, which were in any case administered to the children not as tests of our hypothesis but because of the possible interest of any findings. (Our hypothesis would predict no effect for such a verbal test.) To have included the results (which suggest a mildly positive effect) in support of our theory would have been quite inappropriate.

We also had the chance to administer the Raven Matrices test, but only once, so that there is no test-retest comparison; it was also administered after 4 weeks of supplementation, rather than the 12 weeks in our main study. The insignificant overall superiority of the supplementation groups replicates the finding by Nelson *et al.*² that short periods of

supplementation are insufficient to produce marked changes; this too is not relevant to our main hypothesis, and cannot be used to disprove it. But even here, for reasons given in our article, the largest dose does give a significant degree of superiority over the control group.

Finally, the MAT group test showed the same overall pattern of results (best for 100 per cent of the Recommended Daily Allowance, worst for the placebo) on the WISC-R, but the large standard deviation that we found suggests that the statistical power of the test is insufficient, perhaps because of the children's lack of interest in the test often observed in group testing. This suggests that for replication of future studies, individual tests might be preferable.

Yet Blinkhorn suggests that, to prove our case, we used "several forms of analysis chosen to suit a preferred hypothesis as proof of that hypothesis"; this is difficult to understand. We used the same method of analysis for the MAT and for the WISC-R data, and reported nonsignificance. To find a result nonsignificant does not seem to suggest that we picked arbitrarily a type of analysis which would make the results appear significant. To call this "a blatant piece of post-hocery" seems hardly the measured kind of comment one expects from a fellow scientist.

Blinkhorn argues that, because the results of the blood sample analysis were not yet completed, we should have delayed publication. But the blood analysis data relate to a different problem from that treated in our report, that of *which* vitamin or minerals are responsible for the effect established in our study. Similarly, the data from our British study are a replication of our US study; they will strengthen the conclusions, but we believe that the published data are sufficient to establish the conclusions.

Some of the demands for further information by Blinkhorn and others are being met by S. J. Schoenthaler³, but one point may need comment. Blinkhorn asks: "What of those children who did worse than expected in the treatment group?" It is difficult to understand the question, but the implication that the treatment might have lowered their IQ is breathtakingly naive. Obviously there will be a chance distribution about the mean, so that some do better, others worse than the average. There is no hint in the data of any significant deterioration of IQ in any of the supplementation children; Blinkhorn is welcome to re-analyse our data to satisfy himself on that point.

Blinkhorn finally criticized me for saying that the best policy would be to administer a 100 per cent RDA dose to all children. He leaves out the context, which makes it plain that only a minority would benefit, that such a minority could perhaps be identified by blood sample analysis, but that, in the absence of such analysis, general supple-

mentations might be the best policy. He also overlooks our repeated statement that improving a child's diet would be the most useful way of improving his or her supply of vitamins and minerals.

These are only some of the inaccuracies and errors in Blinkhorn's review; readers are invited to read our symposium to make up their own minds.

H. J. EYSENCK

*Institute of Psychiatry,
De Crespigny Park,
Denmark Hill, London SE5 8AF*

1. Schoenthaler, S. J., Amos, S. P., Eysenck, H. J., Peritz, E. & Yudkin, J. *Personality and Individual Differences* 12, 351-362 (1991).

2. Nelson, M., Naismith, D. J., Burley, V., Gatenby, S. & Geddes, N. *Br. J. Nutr.* 64, 13-22 (1990).

3. Schoenthaler, S. J. *Nature* (in the press).

Scientific licence

SIR — In his review of Frank Close's book *Too Hot to Handle: The Race for Cold Fusion* (*Nature* 350, 29; 1991), Sir Brian Pippard ends: "The book should be read as an exemplary tale by all who are concerned about the conflicting demands of scientific integrity, personal ambition and public interest." Other professional groups have faced such conflicting demands by laying down enforceable codes of conduct for their members, for example the medical profession. I suggest the following scheme will go a fair way towards balancing Pippard's conflicting demands.

(1) Scientists must be licensed to practice, on the lines of medical practitioners, by professional watch-dog bodies. Scientists against whom cases of scientific malpractice and/or fraud have been proved should have their licences withdrawn for varying periods of time [depending on the severity of the malpractice] including, in extreme cases, permanent cancellation. No doubt a malpractice insurance market for scientists will develop — no harm in that.

(2) Analogous to the 'polluter pays' principle, scientists who wilfully pollute the scientific environment should be made to pay the costs of 'clean-up', that is, the costs of all the cross-checking, experimentation, conferences and theoretical wild-goose chases that other scientists are induced to engage in, only to discover a near-hoax. Only a fraction of the costs of such clean-up can be borne by the 'polluting' scientists themselves; the major costs will have to be borne by public- and private-sector employers, who will therefore have a strong incentive to enforce high standards of integrity and validation on the scientific output of their employees.

V. SIDDHARTHA

*Secretariat of SA to RM,
Ministry of Defence,
Room No. 131, 'B' Wing,
Sena Bhavan, DHQ PO,
New Delhi 110 011, India*

Due credit

SIR — I was glad to read Steven Dickman's favourable report (*Nature* 350, 179; 1991) on the European Research Conferences organized under the auspices of the European Science Foundation and to be funded, it is hoped, through the European Communities. I should, however, point out that even though I strongly back this promising scheme to bring the most talented young European researchers together in the formative period of their scientific careers, the men who deserve most of the credit for having established this programme are Professor Heinz Georg Wagner, Universität Göttingen, who began to work for this goal more than two years ago when he was a vice president of the Deutsche Forschungsgemeinschaft, and Professor Claude Fréjaques, vice president of the European Science Foundation.

HUBERT MARKL
(President)

*Deutsche Forschungsgemeinschaft,
D-5300 Bonn 2, Germany*

Selfish psycho-darwinism

SIR — For my present purpose, I wish to imagine that Richard Dawkins and I are the sole sources of male chromosomes on a planet which has a large female population and which we both know will be the subject of unpredictable environmental changes over the coming millennia.

We have been given five wishes in respect of the 'evolvability' of the genetic inheritance we are to pass on. My rival, an honourable man, and confident that the selfish gene model cannot be improved upon, declines this offer. I, on the other hand, more machiavellian, and experienced in human selection methods, make the following stipulations:

(1) Each of my heirs must continually assess its performance in relation to a reference group made up of similar individuals, in all matters affecting genetic replication.

(2) Where its own assessments and the reactions of its peers reflect a comparatively high level of performance, it will experience a sense of well-being. This in turn will trigger physical responses that will act as a multiplier upon its rate of genetic replication by, for example, stimulating its libido and making it more active generally, further increasing its attractiveness to mates, discouraging predators, warding off diseases and intimidating rivals.

(3) Where its own assessments and the reactions of its peers reflect a comparatively poor performance, it will become depressed. The only means of relieving this depression will be: (i) to raise comparative performance; (ii) to adopt a well-received social role, supportive of the high achiever; (iii) to benefit from an environmental change to which it proves better suited than its peers (the 'Admirable Crichton Effect'); (iv) to break away from the reference group by successfully adopting a novel lifestyle.

(4) Where none of the above options is practical, its emotional state will trigger physical changes that will serve to eliminate it from the genetic stock by, for example, suppressing its libido, deterring mates, attracting predators, reducing resistance to disease and making rivals contemptuous.

(5) Where, as a result of option 3(ii), increasingly complex social structures start to develop, genetic factors will start to produce different personality types in order to provide suitable candidates for each role.

A number of questions arise:

(a) Although the above is a simplification of an idea I have been developing for a decade, is it not clear that my 'evoholics' would out-compete Dawkins' heirs, leading to the latter's rapid extinction?

(b) Would they not similarly deal with any of their own kind who, through genetic variability, attempted to backslide to Dawkins' more *laissez-faire* regime?

(c) Have not various research findings over recent years shown that all the 'gifts' I have hypothetically given to my heirs actually

exist, at least among some species? (See, for example, the 'Duke of Marlborough Effect', described on page 286 of the 1989 edition of *The Selfish Gene*. Other references can be supplied on request.)

(d) Does not this hypothesis elegantly synthesize the group selection and selfish gene models? After all, it is an individual whose genetic interests are optimized: Mr (or Mrs) Primogenitor. However, its out-workings would in some circumstances appear to show group selection at work, while in others seem to be totally focused upon the individuals currently bearing Primogenitor's genes.

(e) Finally, does it not answer, more fully than he was at that time able to, the following

Cholera mystery

SIR — I was surprised to see that you had reported that "Dr John Wood . . . removed the pump-handle from a polluted communal water supply in the City of London's Broad Street, bringing a cholera outbreak quickly to an end" (*Nature* 350, 640; 1991). Until now, I had always thought it was Dr John Snow who had removed the handle from the Broad Street pump. On examining several reliable references¹⁻⁹, however, I have failed to find an actual statement that Dr Snow removed the handle. Thus, as sometimes happens in research, an assumption accepted as truth may be erroneous. Who actually did remove the handle?

REBECCA C. GANTT

Philip Morris USA,
120 Park Avenue,
New York, New York 10017, USA

1. Mausner, J. S., & Bahn, A. K. *Epidemiology. An Introductory Text* 93, 94 (Saunders, Philadelphia, 1974).
2. Fox, J. P., Hall, C. E. & Elveback, L. R. *Epidemiology, Man and Disease*, 2, 24-26, 28, 116, 167, 240, 284 (Macmillan, London, 1970).
3. Macmahon, B., & Pugh, T. F. *Epidemiology, Principles and Methods* 7-10 (Little Brown, Boston, 1970).
4. Lilienfeld, A. M. *Foundations of Epidemiology* 24, 25 (Oxford University Press, New York, 1976).
5. Last, J. M. (ed.) *A Dictionary of Epidemiology* 2nd Edn, 122 (Oxford University Press, New York, 1988).
6. Holland, W. W., Detels, R., Knox, G. & Breeze, E. (eds) *Oxford Textbook of Public Health*, Vols 1-IV (Oxford University Press, 1986).
7. Sartwell, P. E. (ed.) *Maxcy-Rosenau Preventive Medicine and Public Health* 10th Edn, 203 (Appleton-Century-Crofts, New York 1973).
8. Pelling, M. *Cholera, Fever and English Medicine 1825-1865* Oxford Historical Monographs, 223-225 (Oxford University Press, 1978).
9. Benton, H. H. in *The New Encyclopedia Britannica* Vol. IX 302, 582 (Chicago, 1974).

SIR — The London physician involved in the cholera outbreak in the 1850s was Dr John Snow not Dr John Wood. More information about his life history and work can be obtained from any reputable epidemiology text. The second error in your report is that he did not bring the cholera outbreak to an end by removing the pump handle as is generally thought. In fact, the epidemic was in its declining phase at that point. His action does, however, stress the need for epidemio-

question raised by J. S. Price in the 29 July 1967 edition of *The Lancet*:

States of excessive depression, anxiety, and irritability are very common. Possibly one in seven of the population consult their general practitioners every year for some emotional disturbance. Why should this be so? We are the result of a process of natural selection, the length and ruthlessness of which confound the imagination. It is known that the severer mental disorders are associated with a reduced fertility, and it is more than likely that under primitive conditions even the milder states of depression conferred a disadvantage in the struggle for survival. Why then are we as a species lumbered with these most disagreeable tendencies — why are we all not paragons of calm, energetic happiness?

M. J. C. WALLER

56 Tennyson Road,
Harpenden, Herts AL5 4BB, UK

logists to take action rapidly; it has also led to one of the definitions of an epidemiologist, namely "someone who sails to glory on the declining phase of an epidemic".

DEREK YACH

Centre for Epidemiological
Research in Southern Africa,
PO Box 19070, Tygerberg 7505,
South Africa

Help wanted

SIR — During his 18 years of working life, Alan Dower Blumlein (1903-42) was awarded 128 patents for inventions which improved almost every aspect of telephony, sound recording, television and radar, and included in 1931 his Specification 394,325 for stereophony.

I am working on a biography of Blumlein and would be grateful if any of your readers could send me copies of documents they may have bearing on his work, including copies of published papers. Please mark the envelope 'BB' to arrive before mid-August 1991 if possible.

F. P. THOMSON

39 Church Road,
Watford, Herts WD1 3PY, UK

All over now

SIR — In his review of the book *Too Hot to Handle: The Race for Cold Fusion* by Frank Close (*Nature* 350, 29; 1991), Sir Brian Pippard concludes that the initial euphoria having now faded, the episode is "hardly more than a footnote to history", and that the institution of science is still robust. I should like to add a modern cleriehew as a codicil to his comments:

*The cheers for Cold Fusion
Were last year's illusion:
What's left of a quorum
Is the Pons Asinorum.*

HARMON CRAIG

Scripps Institution of Oceanography,
University of California, San Diego,
La Jolla, California 92093, USA

Keeping the ivory trade banned

Mark Pagel and Ruth Mace

The most effective way to stop the killing of elephants is to ban the ivory trade. All other options keep ivory in the marketplace and thereby encourage demand for ivory and its products.

THE international moratorium on the trade of ivory agreed at the Convention on International Trade in Endangered Species (CITES) meeting in 1989 is under siege. The ban was already somewhat tattered at birth because several key ivory-exporting countries refused to cooperate with it. But it has achieved a greatness of sorts: after eight frenzied years in which poachers reduced the African elephant populations by 50 per cent or more, by all accounts poaching has plummeted along with the price of ivory¹⁻³. Elephant tusks that would have fetched perhaps \$100 per kilogram for a poacher in 1989, currently find few buyers at \$2-3 per kilogram, according to Richard Leakey, director of the Kenya Wildlife Service². Leakey adds that elephants are beginning to return to parts of northern Kenya from which they had been eradicated or fled³. But rumblings from ivory-exporting and importing nations to lift the ban have placed it squarely on the agenda of the next CITES meeting in Japan during January 1992.

It is easy to be cynical about this issue. Pressures to lift the ban derive from its success: ivory-exporting countries are feeling the pinch of having lost their markets, ivory importers are chafing to resume business, and the decline in poaching has created a sense that elephants are, perhaps, not as threatened as once feared. Moreover, arguments against bans seem to be fashionable in the environmental economics and development literature, in academic enclaves, among science journalists, and even among some conservation organizations^{2,4,5-7}. Politicians have been quick to seize this momentum. The British Trade Secretary David Heathcoat-Amory, following a short debate in the House of Commons last August, said "there are indications that the ban on the trade of ivory is proving effective. . . I hope that should the situation continue to improve it may be possible to in the not too distant future allow the ban to be lifted in some respects" (ref. 1). But despite the fall in ivory prices, poaching is still occurring.

More thoughtful attempts to argue against

the ban have been mooted by the economists^{4,6,7}. An indefinite ban on the trade in ivory, the argument goes, causes an initial drop in demand and price as legal demand for ivory declines. But due to latent consumer demand for ivory, some of the formerly legal demand will be shifted into illegal demand (that is, demand that can only be satisfied by illegal supply). This argument assumes, probably correctly, that any ban will

world of ivory poachers slaked by illegal consumer demand for ivory is distressing. Nevertheless, the economic arguments hide the gruesome fact that nearly all of the so-called legal supply of ivory has in the past been obtained illegally by poaching. Only two countries — Zimbabwe and South Africa — have legal culling programmes and they provide only a small fraction of the total ivory that is traded⁵. The remainder of the

Ardea ivory has been confiscated from poachers by governments and sold. This point is overlooked by every economic argument advanced against the ban, and yet it is crucial to understanding why lifting the ban must be avoided. Governments that confiscated poached ivory were, under CITES pre-ban Appendix II regulations, able to sell this ivory legally. Appendix II regulations allow some legal trade in ivory whereas the more restrictive Appendix I regulations — in force under the current ban — prohibit trade in elephant products of any kind. Therefore, countries whose elephant populations were being drastically reduced by poaching were still selling 'legal' ivory. Furthermore, given such rules, there was great temptation for government 'confiscators' to team up with 'poachers', leading in some cases to the ivory being 'confiscated' directly from the elephant.

When weighing arguments against the ban, it must also be borne in mind that total demand for ivory is lower if there

African elephant — worth more alive than dead?

be cheated upon because some countries will ignore it either by allowing the killing of elephants for ivory, or by allowing the sale of ivory products. The result of increasing the demand on a subset of the total supply (that is, on the supply that is obtained by poaching) is to raise the price of ivory, and thereby promote poaching. Over time the amount of illegal demand may even increase if more and more of the formerly legal demand shifts to illegal demand, or if new consumers enter the market. The result is even more poaching. Lifting the ban, it is argued, would prevent this upward spiral of poaching.

The spectre of a brutish hobblesian under-

is at least one honest ivory trader who does not respond to the ban by seeking illegal supplies, or at least one obedient consumer who decides not to buy ivory products. So, if the goal is fewer elephants killed, a ban will probably work. The success of the current ban attests to this. But those opposed to a ban argue that a potential difficulty arises because poaching is not easily stopped. Regardless of how one feels about killing elephants, there is some amount of ivory that could be obtained each year that would be sustainable over the long term. If enough countries were to ignore the ban, especially the importing nations which create the de-

mand for ivory, poaching might exceed this. Again, however, such arguments ignore the fact that the proportion of the total world demand for ivory that was being met before the ban by truly legal culling was very small. Concern about the amount of poaching under a ban misses this important point.

Other arguments against the ivory ban rely on the idea that a legal trade in ivory gives governments an incentive to keep poaching at a minimum as a way of protecting a valuable export commodity⁴⁻⁷. Economists turned conservationists, the environmental economists, argue that elephants, like other environmental resources, can be "sustainably managed" or "sustainably developed" once they are properly valued⁸⁻⁹. The idea of sustainable development is that a resource has some estimable value to society and to future generations. Using or consuming the resource in a sustainable way can mean a variety of things, but the key point seems to be that the resource is preserved for future generations. For example, if the resource is clean water, then one might want to tax polluters as a way of either giving them an incentive to pollute less, or as a way of accumulating money which is spent on cleaning the water. If the resource is consumed but is renewable, like trees, the idea is to charge a high enough price for them to ensure that there is roughly a balance between consumption of trees and planting of new trees so that the resource will not be used up. In theory, the price attached to the resource should be sufficient to guarantee that the resource is properly managed.

Sustainable development

These ideas are now being applied to elephants^{6,7}. The argument is that if the revenues from sustainable development of elephants, that is from culling or hunting programmes, are higher than those which can be obtained from short-term eradication, there will be an incentive to conserve the elephants. But such schemes can never work for elephants so long as they are a 'commons' or nonexcludable resource. If an ivory trade is resumed in the form of some countries attempting to run sustainable development programmes, the consumer-driven demand for ivory will return. This will be true even if only a small number of countries are given the right to trade in ivory. The demand created by the legal sale of ivory will provide incentives to poachers. If poachers can kill elephants and sell their ivory, the short-term payoffs of doing so will be much more valuable to them than the idea of preserving the elephant for possible long-term advantages to society — the 'intergenerational equity' of the green economists⁹. Only if all elephants — not just those in one or a few countries — could be rounded up and put in a pen would such a sustainable development programme have a chance of working. Then the death of each elephant, perhaps from government culling for later sale or from a tourist hunter's rifle could be 'appropriately' valued. But,

like air, large migratory animals tend to be found all over.

Countries that want to preserve their elephants but still engage in a legal ivory trade will always suffer the costs of stopping the illegal killing of elephants for their ivory. No matter that only some ivory would be 'legal': smugglers are notoriously good at getting around legal strictures, and certificates of legality can and have been forged^{4,10}. Further, attempts by governments to stop poaching will always have to bridge a fundamental asymmetry between poachers and governments. Poaching is a livelihood for many poachers. Stopping poaching is a policy that can return benefits, but not ones on which the livelihood of the government rests. There are analogies here to the state of nature: a rabbit's life depends upon not getting caught by the fox chasing it. Not catching a rabbit means the fox may go without its dinner. Running fast brings higher rewards to rabbits than to foxes¹¹. Aesop knew this.

Managing elephants for long-term yields is also difficult to sustain because of the presence of other nations that want to make a profit from their ivory. If one or two countries were to be granted trading status, and they were to make money from ivory, it would only encourage other nations to do the same. Managing for long-term sustained yields would require cartel-like cooperation among these nations to prevent increased ivory production by one or more of them. Where the attraction of short-term economic gains — often to ameliorate domestic ructions — can easily outweigh the more abstract promise of sustainable but lower long-term yields, this is unlikely: cartels are unstable, just look to OPEC.

The most effective way to stop the killing of elephants is to reduce demand for ivory. This is best achieved by banning the trade in ivory and making social outcasts of those in possession of it; strategies that seem to be working for furs. Allowing even a small legal trade in ivory will make this impossible: the existence of any socially acceptable ivory would confuse public opinion, and keep consumer demand alive. The reduction in demand that has been brought about by the publicity of the current ban has probably been the most important reason behind its success in reducing poaching⁴. Bans or partial bans enacted by Britain, the nations of the European Community, the United States, Dubai and Japan may all have been a response to public pressure. A ban is the only option that avoids public confusion and which greatly reduces the demand for ivory, even if some countries ignore it. It is the much more trying option: a ban will create social disruption as ivory carvers lose their market, and it will witness the temporary degradation of habitats as some elephant populations grow large. But elephants have proven to be a great attraction to tourists — thus generating jobs, and revenues — and nature has long known how to regulate natural populations. Moreover, elephants as a tour-

ist attraction do not require the elephant-containing nations to cooperate like a cartel: there are no incentives to cheat on the elephants. Just the opposite behaviour — competition for tourists — will be favoured.

Tourism

Tourism also provides a dependable incentive for stopping poaching. In addition to the money that tourism funnels into local economies, tourists may even be willing to pay an elephant surcharge⁵. If elephants are worth more alive than dead then they will not be killed. This is different from the sustainable development idea, which attempts only to value appropriately each kill, and which by placing ivory in the marketplace keeps demand alive. Kenya is already profiting handsomely from showing off its elephants to tourists. The key to the success of tourist programmes such as Kenya's is to direct the profits just as much as possible into improving the standard of living of the local people involved in the tourist trade, and into improving the management and facilities of national parks.

Just such a programme has been highly successful at providing protection for the mountain gorillas in Zaire¹². Poaching had reduced the population to around 260 but now the ranks are increasing. Receipts from allowing groups of just six tourists at a time to visit gorilla families rose from zero to about £23,000 per month between 1984 and 1987 (ref. 12). Guards are provided with uniforms, boots, warm clothing, sleeping bags and wet-proofs¹².

Any other option for the ivory trade, such as relying upon centralized trading exchanges^{2,6,7} as a way of monitoring where and how the ivory was obtained (for example, by using isotope analysis of tusks^{13,14}), and to whom it will be sold, will only encourage the demand for ivory products and suffer from the weaknesses discussed here. Such schemes reflect economic and market ideals, fanned by current scholarly fashion, more than reality in an industry that has proven adept at getting its product (legal and illegal) to market so long as there is demand for it. □

Mark Pagel is in the Department of Anthropology, Harvard University, Cambridge, Massachusetts 02138, USA; Ruth Mace is in the School of Development Studies, University of East Anglia, Norwich NR4 7TJ, UK.

1. *New Scientist* **127**, 26 (1990).
2. *The Economist* **318**(7696), 16 (1991).
3. *The Economist* **318**(7696), 42 (1991).
4. *The Economist* **312**(7609), 15 (1989).
5. *The Economist* **313**(7624), 19 (1989).
6. Barbier, E. & Swanson, T. *New Scientist*, **127**, 52 (1990).
7. Barbier, E., Burgess, J., Swanson, T. & Pearce, D. *Elephants, Economics, and Ivory* (Earthscan, London, 1991).
8. *The Economist* **312**(7620), 77 (1989).
9. Pearce, D., Markyanda, A. & Barbier, E. *Blueprint for a Green Economy* (Earthscan, London, 1991).
10. Milner-Gulland, E. J. & Mace, R. *Biol. Conserv.* **55**, 215–229 (1991).
11. Dawkins, R. & Krebs, J. R. *Proc. R. Soc. B* **205**, 489–511 (1979).
12. Aveline, C. & Aveline, R. *Oryx* **23**, 64–70 (1989).
13. Vogel, J. C., Eglington, B. & Auret, J. M. *Nature* **346**, 747–749 (1990).
14. van de Merwe, N. J. et al. *Nature* **346**, 744–746 (1990).

Contaminated origin of AIDS viruses

That Gallo's AIDS virus may have been contaminated with the French strain has always been on the cards, but now it seems that even the original French virus was contaminated. Will this put an end to past squabbles?

At the end of February, Robert C. Gallo of the US National Cancer Institute dropped a bombshell when he reported (*Nature* 349, 745; 1991) that the French AIDS virus LAV, said for years to have come from a Parisian fashion designer known as BRU, was not from BRU after all. Gallo could not say where LAV did originate, but by sequencing virus from a sample that Luc Montagnier of the Institut Pasteur sent him in September 1983, taken from BRU, Gallo's colleagues Hong-Guang Guo, Marvin Reitz and others showed that this authentic or true BRU virus differs from the LAV sequence the Pasteur published in 1985. LAV, said Gallo, is LAV, but it is not BRU. But if LAV is not BRU, Gallo's own principal human immunodeficiency virus (IIIB) used for the blood test his laboratory developed in 1984, could not have come from the Pasteur.

Not so fast. After reading Gallo's letter, Gerald Myers, who heads the AIDS virus database at the Los Alamos National Laboratory in New Mexico, challenged the data, arguing that the criteria used to estimate the genetic distance between the viruses were flawed. Where Gallo and Reitz reported that true BRU and LAV/IIIB differed by 8 per cent, Myers said that they had not followed the accepted convention of not counting gaps in viral sequences. By his calculation, true BRU and LAV/IIIB are only about 4 per cent different and, therefore, are in a twilight zone.

Now, it turns out that Gallo was right about BRU, the confirmation of his analysis coming, ironically enough, from Montagnier himself, who reported with Simon Wain-Hobson and others in last week's *Science* (252, 961-965; 1991) that Gallo and his colleagues have identified true BRU. However, the Gallo team agrees that the difference drops from 8 to 4 per cent using Myers' methods (the details will appear next week).

Montagnier discovered for the first time that there was a contamination in his own laboratory, most probably during a two-week period in July 1983 when material from the patient BRU was being examined alongside samples from a couple of other AIDS patients.

The virus known as LAV, Montagnier now reports, really came from a patient known by the letters LAI — like BRU a French homosexual who had been in the United States in the late 1970s, BRU in New York and LAI in the midwest. After insisting for eight years that there had been a contamination in Gallo's laboratory, Montagnier

now writes that "contamination is a recurrent problem in microbiology."

Do these new data, gathered in bitter competition, mean that the dispute between Gallo and Montagnier over the true nature of LAV/IIIB (known as HIV-1) can be settled? One would hope so. Unquantifiable hours of brainpower that should be devoted to curing AIDS have been wasted on the argument whether Gallo's IIIB really derived from the Pasteur. Scientifically it does not matter. Legally it does not matter either; both sides agreed to share credit and divide royalties from the AIDS blood test in 1986. They also agreed not to reopen the disagreement.

Sadly, to think the end is in sight may be just wishful thinking, for there is one strange and ironic element to the story that leaves each side as suspicious of the other as before. It has to do with a sample, originally thought to come from BRU, known as M2T-B.

As is now well known, the French sent supernatant to Bethesda in July 1983 and, when attempts to grow virus from that proved fruitless, sent two further samples in September — labelled JBB/LAV and M2T-B/LAV. Each sample was said to have come from BRU.

JBB/LAV would not grow in established T-cell lines but M2T-B/LAV showed promise, as did a couple of other samples, notably those known as RF and MN. Still, Gallo's colleague Mikulas Popovic was having no success in getting virus growth in quantities sufficient for a blood test. So in an experiment that many consider so bizarre that, as one Pasteur researcher says, "even a drunk wouldn't do it", Popovic pooled samples from ten different AIDS patients. His hope was that there would be enough reverse transcriptase in the pool to stimulate one of the ten putative AIDS viruses to grow.

It worked. Popovic's virus came to be called IIIB. The blood test that is now the mostly widely used around the world followed.

So did a bitter controversy, when it was subsequently discovered that the sequence of IIIB is virtually identical to the sequence of Montagnier's LAV. Popovic has stated repeatedly that LAV was not among the ten samples in the pool. Nevertheless, most people have concluded that LAV must have contaminated his material and reemerged with a new name.

Gallo suggested at the time that a mix-up might have occurred one night when, at short notice, his laboratory moved to new quarters. Still, he has never until now fully accepted the idea that IIIB is really LAV.

Why? Pride? Competitiveness? Chauvinism? Perhaps. But underlying all that was a persistent belief that the two viruses might not be the same because biologically they do not behave the same, even though their nearly identical structures suggest they should. Simply put, JBB/LAV from the patient BRU does not grow in T cells; IIIB does.

Gallo saw this as proof that IIIB could not have come from the samples he received from Montagnier. But there was one catch. The sample M2T-B had long since been used up, and so could not be analysed. Nevertheless, as the French attested to its provenance from BRU, Gallo assumed that it, too, would look like JBB — now known as authentic or true BRU.

In any case, it seemed to explain the question of biology. JBB/LAV, now JBB/BRU, does not behave like LAV/IIIB because it is a recognizably different virus.

The irony is that the French discovered LAI in one of their original samples of M2T-B. It is now clear that a contamination occurred.

Why does this not end it? When contamination is proved to have occurred on both sides of the Atlantic, which group can claim to be scientifically cleaner than the other?

The problem is M2T-B. The Americans say they find it odd that the origin of LAV should turn up in the one sample they had just reported they no longer had. The French say darkly that they find it odd that M2T-B is the one sample the Americans no longer have.

Meanwhile, the new findings raise at least one question that needs pursuing: double infection. When Gallo reported finding BRU but not LAV in the JBB sample, the possibility of double infection was raised. Was Gallo sure LAV was not lurking undetected? Perhaps another test should be done. Because there have been no reports in the literature of double infection, the notion was dismissed.

But the new data make it plain that the M2T-B sample was doubly infected. It is not yet known whether the patient LAI harboured two viruses, but Wain-Hobson will pursue the issue. Meanwhile, in another report in this issue of *Nature* (page 277), Jean-Claude Chermann, who was at the Pasteur in 1983 and 1984, reports what appears to be a double infection of his 1984 BRU samples. The jury is out, but if it is shown that people can be infected with more than one strain of HIV, it will not help vaccine development.

Barbara J. Culliton

Of size and abundance

John Damuth

Body size has long been thought to influence both the distribution and abundance of animal species, but we are only just beginning to discern the overall role that size plays in organizing the Earth's biota. On page 312 of this issue¹, Nee *et al.* clarify aspects of the picture, reporting the results of a new study of British birds that reveals remarkable similarities between the way that regional and local abundances relate to body size. Particularly striking is the finding that, at certain scales of analysis, larger body size does not mean fewer individuals, when the general trend is the reverse.

Nee *et al.* describe the relationship between body size and abundance in a total sample of 147 bird species and then separately within the lower taxa represented in the data set. The data come from the censuses of the British Trust for Ornithology. These are virtually unique in consisting of comparable estimates, for a large sample of species, of total numbers of (breeding) individuals over a significant geographical area — in this case, all of Britain. Nee and colleagues show that, overall, larger species are represented by fewer individuals, abundance decreasing as the -0.75 power of body mass. In contrast, abundance tends to increase with body mass within lower taxa such as genera and tribes.

Surprisingly, the second trend is most pronounced among those tribes that have no close relatives in Britain. Why this should make a difference is not exactly clear, but Nee *et al.* reason that phylogenetically dis-

tinct tribes are probably ecologically distinct as well. This further implies that such a tribe is more likely to comprise a complete guild — all the species present that play a particular ecological role. In contrast, tribes closely related to other tribes in Britain probably more often represent fragments of guilds that have members in more than one tribe. Thus Nee *et al.* suggest that the positive relationship between size and abundance is fundamentally a characteristic of avian guilds. The same patterns are found among Swedish birds, although the data are not as good and there are some differences in the patterns within taxa.

In spite of a growing literature, it has been difficult for researchers to form a consensus about general patterns of size–abundance scaling. Initially, many assumed that the same patterns should be observed over all scales of analysis. The apparently conflicting results from studies using different types of data, and carried out across different ecological, taxonomic and geographical scales, have left the field in some disarray². Many of the inconsistencies can be resolved by recognizing that different analyses provide insights into different biological questions, and enough information is now available to perceive some order in the apparent confusion — indeed, when the new results are compared with those from other analyses of size and abundance some intriguing regularities begin to emerge.

The most thoroughly studied relationships between size and abundance are those involving what are called 'ecological' population densities. Unlike regional densities, an ecological density is measured only in an area that is actually occupied by the species. Ecological density thus expresses a species' effect on its local resources and the degree to which it has successfully adapted to extract energy from its local environment. A number of studies spanning many taxa and a large size-range of species have shown that ecological density decreases approximately as the -0.75 power of body mass^{3–6} (Fig. 1). Because individual metabolic requirements scale reciprocally, as the $+0.75$ power of body mass, this means that the amount of energy each species uses per unit area of its habitat is independent of body mass. Nee *et al.* refer to this as the 'energetic equivalence rule'.

What is going on in the relatively narrow region of body size encompassed by birds (Fig. 1) has been subject to debate^{5,7}. A weakly negative relationship for all the birds is usually found across guilds, but one would not expect the overall relationship to be clearly reflected in such a size-restricted subset of the biota. What is clear is that birds are not really anomalous, in that ecological densities reported for them are for the most part

consistent with the energetic equivalence rule across terrestrial species.

However, when ecological densities are compared within guilds or groups of closely related small species of birds or mammals, a different pattern is found. Frequently, the slopes on double-logarithmic plots (exponents) are less steep than -0.75 , showing that larger species are better at getting and using energy than are smaller members of the group⁸. An exception might be among guilds

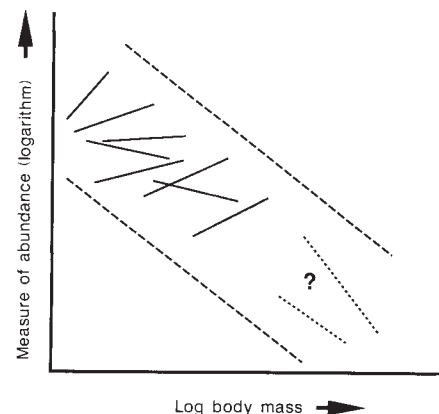


FIG. 2 Diagram of idealized empirical relationships between abundance and body size at different scales of analysis. For both ecological and regional ('crude') densities, the overall relationship across a diverse array of species (within the region between dashed lines) is negative and usually consistent with the energetic equivalence rule (slope about -0.75). But within guilds or groups of closely related small species in the overall data set (solid lines), both locally and regionally the slopes tend to be shallower than -0.75 or even positive, showing that the larger guild members usually control more energy. Among feeding guilds of larger mammals (dotted lines), however, guild slopes may often be equal to or more steeply negative than -0.75 .

that include medium-sized to large mammals, where steeper slopes than -0.75 have been observed⁹.

A comparison of ecological densities, such as those shown in Fig. 1, tells us nothing however about distribution. A species might be common where it lives, but that particular habitat may be very restricted. Likewise, a locally rare species might be distributed widely. Regional densities (also called 'crude' densities), which express the total number of individuals of each species in the region, could thus exhibit notably different patterns to those observed for ecological densities. Differences in distributions and total population sizes of species could have important consequences for, among other things, their conservation and macro-evolutionary fate^{10,11}.

Unfortunately, a plot corresponding to Fig. 1 for regional densities cannot currently be drawn with any confidence, because such densities are unavailable for most taxa other than birds (which are relatively conspicuous and whose habitat requirements are better known). Nee *et al.* provide the best comparison between regional and ecological den-

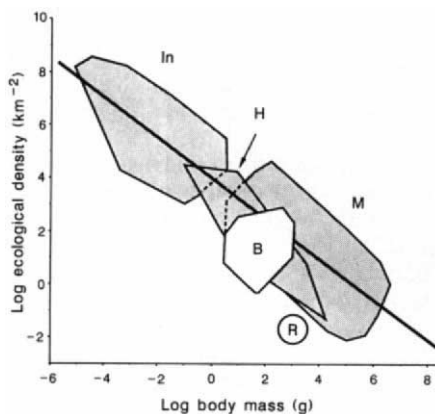


FIG. 1 The scaling of ecological densities (see text) with body mass for terrestrial organisms. Polygons enclose the dispersion of points for each of several groups: B, birds; H, reptiles and amphibians; In, invertebrates; M, mammals; R, nine outlying points for raptors separated from the 555 other densities reported in ref. 5. Density values for poikilothermic ectotherms are adjusted to correct for differences in average metabolic rates from homeothermic endotherms (see ref. 4). The line is fitted through the non-avian data and has a slope of -0.76 . Bird data from ref. 5, other data from ref. 4.

sities that we can make at present. Their slope of -0.75 for regional densities implies that there is no overall relationship of size, energy use and distribution. Regionally, the total energy used by each bird species is independent of body size, conforming to the energetic equivalence rule. Within guilds, however, larger species use more energy.

The scaling patterns for regional and ecological densities are thus remarkably congruent. We can tentatively offer a single description of size scaling for both kinds of abundance measures as a sort of working hypothesis (Fig. 2).

As yet there is no satisfying explanation for the empirical patterns shown in Fig. 2, but their meaning is clear. From the point of view of energy control, the answer to the question "Is it better to be a small animal or a large one?" may be that it doesn't matter. But the answer to the question, "Is it better to be a small or a large member of your guild (or genus)?" may very well be "large". Unless, perhaps, you are in a large-mammal guild.

Van Valen¹² once found that the total

amount of energy controlled by a species of Caribbean palm was roughly the same as the trophic energy used by all of the individuals of *Homo sapiens*. This result, which remains a curiosity at the moment, nevertheless hints at there being as yet undiscovered patterns involving size, energy, distribution and abundance at even higher scales of analysis. □

John Damuth is in the Department of Biological Sciences, University of California, Santa Barbara, California 93106, USA.

1. Nee, S., Read, A. F., Greenwood, J. J. D. & Harvey, P. H. *Nature* **351**, 312–313 (1991).
2. Lawton, J. H. *Oikos* **55**, 429–434 (1989).
3. Damuth, J. *Nature* **290**, 699–700 (1981).
4. Damuth, J. *Biol. J. Linn. Soc.* **31**, 193–246 (1987).
5. Juanes, F. *Am. Nat.* **128**, 921–929 (1986).
6. Marquet, P. A., Navarette, S. A. & Castilla, J. C. *Science* **250**, 1125–1127 (1990).
7. Brown, J. H. & Maurer, B. A. *Am. Nat.* **130**, 1–17 (1987).
8. Brown, J. H. & Maurer, B. A. *Nature* **324**, 248–250 (1986).
9. Robinson, J. G. & Redford, K. H. *Am. Nat.* **128**, 665–680 (1986).
10. Brown, J. H. & Maurer, B. A. *Science* **243**, 1145–1150 (1989).
11. Arita, H. T., Robinson, J. G. & Redford, K. H. *Conservation Biol.* **4**, 181–192 (1990).
12. Van Valen, L. M. *Biotropica* **7**, 260–269 (1975).

GENERAL RELATIVITY

Evading the cosmic censor

John L. Friedman

A COSMIC censor appears to protect general relativists from one of the more embarrassing consequences of Einstein's classical (that is, non-quantum) theory of gravity. But numerical simulations by Stuart Shapiro and Saul Teukolsky (*Phys. Rev. Lett.* **66**, 994–997; 1991) challenge the censor's iron grip.

Classical gravity is a nearly perfect theory. With pythagorean simplicity, its precise mathematics unites geometry and astronomy. Its predictions of planetary orbits, already remarkably accurate when only the newtonian limit was known, are now verified to one part in 10^9 ; its more exotic prediction of gravitational waves is confirmed to better than one per cent, the result of our luck in finding a gravitational wave detector in the form of a binary pulsar — a pair of orbiting neutron stars that spiral towards each other as they radiate energy in gravitational waves. And isolated black holes probably mimic the ideal geometries of the theory to better than 1 part in 10^{25} .

The flaw in classical gravity is that the histories it predicts are singular: within the theory, stars that collapse fall inward until their densities become infinite; and the Universe itself, run backwards to the Big Bang, was at infinite density only a finite time ago. The theory is in this way incomplete. It predicts the formation, within finite time, of regions of space in which it cannot be valid.

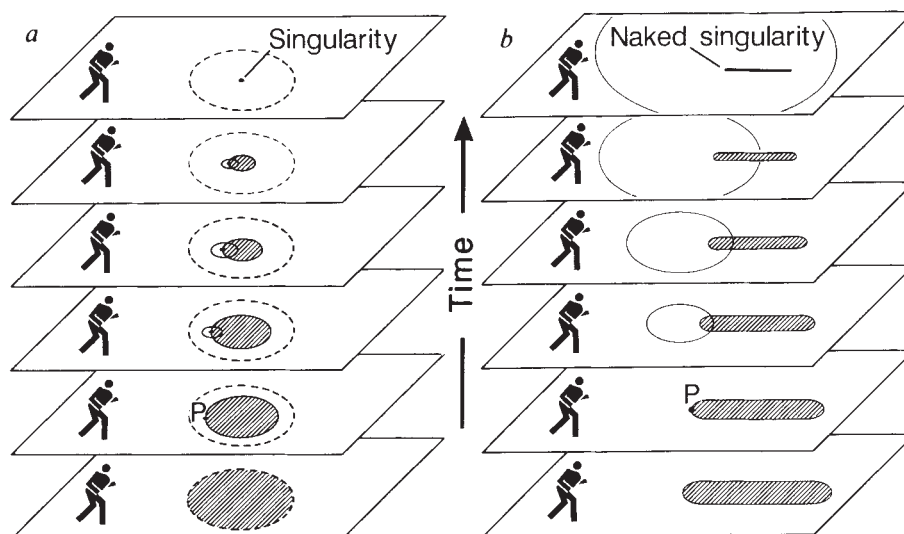
Still, general relativity bears its imperfection with a remarkable grace. Gravity seems invariably to veil its singular behaviour behind the 'event horizons' that form the boundaries of black holes. Only to observers

inside black holes or present at the Big Bang is the singularity revealed, and even then perhaps only when they are themselves crushed to infinite density.

This is the cosmic censorship hypothesis of Roger Penrose (*Riv. Nuovo Cimento, Serie I*, **1**, Numero Speciale, 252–276; 1969), but

in their new work Shapiro and Teukolsky question whether gravity's modesty is perfect. From a calculation that models the collapse of a cluster of particles, they find naked singularities, gravitational collapses without black holes. Whether a more realistic version of the collapses they study would still be singular is open to question. At a minimum, however, their work emphasizes how difficult it is to prove the existence of a cosmic censor.

Penrose's conjecture can be stated in a precise form for a vacuum spacetime. Even without matter, space need not be flat, and gravitational waves — waves of curvature that travel at the speed of light — can collapse to form black holes. The Einstein equations describe an evolution of geometry in which these waves attract each other. For low, long waves the attraction is slight, and they pass freely through one another, eventually to spread like water ripples. Higher waves with shorter wavelengths have higher curvature, and the attraction of the waves counters their spreading. If the waves are sufficiently high or steep, their self attraction stops them from spreading at all, and as they crowd together, they become steeper. Because greater curvature implies greater gravitational attraction, the crowding accelerates, leading to a curvature that grows without bound. The famous singularity theorems of Penrose, Stephen Hawking and others prove that gravity is, in this way, "an essentially unstable force" (*Proc. R. Soc. A* **314**, 529; 1970). Once a black hole forms — once the curvature is high enough to trap light and gravitational waves — the evolution of spacetime will be singular. ►



Black holes and naked singularities. *a*, Gravitational collapse of a nearly spherical collapsing star (shaded) leads to the formation of a black hole, as shown here in a sequence of two-dimensional pictures. Just before a singularity forms, a flash of light is emitted at a point P in the star, and it initially expands outward, forming a sphere (a solid circle in each figure) about the matter that emitted it. Ultimately, however, the light contracts, collapsing, like the star to a singularity. It never penetrates the event horizon (dashed circle) and so never reaching an external observer. *b*, For models for matter with low enough pressure, naked singularities can arise from the collapse of a cylinder to a line. Because the matter is concentrated along only one dimension, its gravity is weaker, and light is not trapped. Light emitted at any time from the collapsing matter will reach an external observer, who therefore sees a singularity form.

Radon activity

EXPOSURE to large quantities of radioactive radon gas in the home is associated with a high frequency of mutations in a gene in peripheral blood T-lymphocytes. This result, reported by B.A. Bridges *et al.* in *The Lancet* (**337**, 1187–1189; 1991), is consistent with the controversial idea that leukaemia is more prevalent in countries where the average indoor concentration of radon is high. The blood samples were taken from 20 non-smoking adults from the village of Street, in Somerset, UK, where many homes receive more than the recommended maximum exposure to radon. The authors point out the tentative nature of the finding, which could be accounted for by chance or some confounding factor.

Deep thought

ANTINEUTRINOS radiated from nuclear reactions in the Earth's core could reveal the chemical composition of the deep interior, argue M. Kobayashi and Y. Fukao (*Geophys. Res. Lett.* **18**, 633–636; 1991). The antineutrinos of interest come from the β -decay of ^{40}K . The problem is in detecting and identifying them. Twenty seven tons of liquid ^3He could help, conclude the author. Collisions of the antineutrinos with the isotope's nuclei produce positrons which radiate scintillation light; with this mammoth detector, one antineutrino may be seen every day. The experiment resembles the neutrino detectors that monitor solar nuclear reactions — hence the article's title, "The Earth as an Antineutrino Star". But enthusiasts should recall that for 20 years the solar results have more perplexed than enlightened.

Spit and polish

OUR saliva contains many proteins, a quarter of them belonging to a family characterized by basicity and an enormous proline content. D. L. Kauffman *et al.* (*Biochemistry* **30**, 3351–3356; 1991) have now isolated eleven of these from one person (diligent spitting by the senior author) and sequenced six. All or most of them are the products of separate genes, and comparisons with published DNA sequences reveal differences that must be attributed to complex genetic variation. Why we should need so many similar proteins in our saliva is not known, but the conjecture is that the biologically important properties are more composition- than sequence-dependent and permit of abundant mutations. The proteins evidently bind to bacteria and may be responsible for attaching them to the teeth and generating plaque. They also, however, lubricate chewing and thus satisfy dentist and epicure alike.

Cosmic censorship is in part the converse, the statement that if the behaviour of spacetime is singular, a black hole must form. But it is stronger: the spacetime outside a black hole must be smooth, its curvature bounded, and the paths of its gravitational waves (or of light rays) must never end, so long as they remain outside any black hole. Although no proof of cosmic censorship has been found, a number of failed attempts to find naked singularities argue in its favour.

Black holes forming in collisions of gravitational waves are initially distorted and dynamic. The waves that are not caught, however, rapidly disperse, and the black hole they leave behind quickly settles down to a final, stationary geometry, characterized solely by its mass and angular momentum. For a given mass, the final black hole can have a maximum angular momentum, roughly that of a ball whose mass and circumference are those of the black hole and which rotates at the speed of light. The assumption of cosmic censorship imposes stringent restrictions on the possible initial spatial geometries, requiring symmetric geometries to have no more angular momentum for their mass than the final black hole. That is, because the waves carry off mass but not angular momentum, the final state would have too much angular momentum to be a black hole (or a set of black holes) and a naked singularity would be the outcome of collapse.

Serious searches have been made for geometries that satisfy the Einstein equations and which violate the conditions that would allow a final black hole state, but they have encountered no success. The equations clearly conspire to avoid geometries that would, at least in this way, remove from singularities the cover of their black holes.

Does adding matter make it possible to evade the cosmic censor? With the simplest kinds of matter, like light, it is no easier to create naked singularities. A collection of particles, however, is difficult to model in a way that correctly accounts for their interactions at high density. If one models matter as, for example, a perfect fluid or as a collection of noninteracting particles, then even in the newtonian theory (in fact, even without gravity) a smooth initial state can have a singular evolution, with points of infinite density. Fluids without pressure (called 'dust') or with unrealistically low pressure cannot keep themselves apart. Several years ago, Kip Thorne suggested an appealing extension of cosmic censorship to exclude such models of matter: singularities that are already present in the newtonian theory will be regarded as uninteresting, the result of the artificial matter, not of the fundamental instability of gravity. Cosmic censorship then requires that if the matter one uses is nonsingular in newtonian gravity, it cannot give rise to naked singularities in the full, relativistic theory. (A similar, but stronger version of the conjecture was proposed last month by Steven Hawking as a bet between him on one

side and Thorne and John Preskill on the other. Hawking will regard matter as artificial if it forms singularities without any gravity at all. He bets that cosmic censorship is true for any matter less artificial than that: if matter needs gravity to be singular, its singularities will all be hidden by gravity's event horizons.)

The example that Shapiro and Teukolsky consider is a collection of particles which interact only gravitationally. The particles are initially in the shape of a prolate spheroid — a long rounded cylinder — their velocities directed inwards towards its axis. As the particles approach the axis, their density rises and the spacetime curvature — the gravitational field of the particles — increases in tandem. When they reach the axis, their density is infinite, and so is the curvature, but if the spheroid is long enough, no black hole clothes the singularity.

It has long been known that finite cylinders of dust and cylindrical collections of collisionless particles collapse to singularities in the newtonian theory. Thus, by Thorne's criterion, one could say that the naked singularities arising from the collapse of collisionless matter do not violate cosmic censorship.

However, Shapiro and Teukolsky claim that the singularity really extends beyond the collapsing matter, and if it does there is cause for alarm. Outside the matter, the singularity is not simply the result of an infinite density of matter — matter that, by design, is incapable of pushing itself apart. At least it cannot be simply the result of using artificial matter if an observer sees the exterior singularity develop at the same time as the singularity of the matter. For then, even if one changes the matter so that it collides and bounces just before infinite density is reached, it will not be possible to send a signal from the matter to the geometry outside in time to prevent the collapse. A finite speed of light implies that there would be no way to prevent the curvature outside from becoming infinite (see figure).

In fact, if Shapiro and Teukolsky are right, one ought to be able to replace their collisionless particles by a collection of small black holes, and to violate in this way the conjecture that seemed most secure — cosmic censorship for vacuum spacetimes. The smaller the black holes, the less often they would collide. One could approximate with arbitrary accuracy the collisionless particles that Shapiro and Teukolsky model, and even spacetimes without matter would have naked singularities.

For now the question remains open. The evidence has more than enough ambiguity to allow prejudice to rule: gravity without matter seems too graceful a theory to allow a breach so crass of a modesty so persistently maintained. □

John L. Friedman is in the Department of Physics, University of Wisconsin, Milwaukee, Wisconsin 53201, USA.

Half of a peptide pump

Peter Parham

ON page 323 of this issue¹, Spies and DeMars provide solid evidence to support the idea that 'transporter' genes in the major histocompatibility complex encode pumps that deliver peptides to antigen-presenting molecules. Although a convincing demonstration of the importance of the gene they looked at, their experiment does not directly address its mechanism of action. Transporter function still remains putative — but less so than before.

Every immunologist knows that proteins must be made presentable for T cells to see them. For viral and endogenous proteins this involves intracellular degradation to an oligopeptide, binding to a class I major histocompatibility complex (MHC) molecule, transport to the cell surface and, finally, engagement with the T-cell receptor. *In vivo*, peptide binding to class I MHC molecules is restricted topologically to the endoplasmic reticulum. On the other hand the source of peptides can be proteins found in a variety of cellular locations including the cytoplasm, nucleus and mitochondrion. That intracellular membranes prevent simple diffusion of either intact or degraded proteins from these compartments to the endoplasmic reticulum, pointed to involvement of a specific transport mechanism². Consistent with this notion was the characterization of 'regulatory' genes within the MHC that affect assembly of class I molecules and their transport to the cell surface^{3,4}. Further raising stock in this hypothesis was discovery, at an appropriate place in the MHC, of genes homologous to those encoding a populous family of ATP-driven membrane transporters⁵⁻⁹.

Although strongly suggestive, this evidence was at heart only a correlation and none of the four groups who cloned the genes could attest to their function. Spies and DeMars now show that transfection with a putative peptide transporter gene cures some — but not all — cells which are defective in the expression of class I MHC molecules.

In principle the experiment was simple, but practice, as intimated by the absence of any gene expression in the previous four

reports⁵⁻⁸, has been another matter. Spies and DeMars initially found transfection with genomic cosmid clones failed and their subsequent success required complementary DNA clones driven by heterologous promoters, which ironically were of viral origin. Perhaps another key to their achievement was an unrivalled choice of recipient cells in which to transfect their gene.

DeMars has exploited genetic polymorphism, the bane of transplant surgeons and perhaps the human genome project, systematically to derive cell lines with heterozygous and homozygous mutations in the MHC. For the human MHC, wild types do not exist and heterozygosity is the norm. Moreover the products of antigenically distinguishable alleles are codominantly expressed at the cell surface. Thus rounds of mutagenesis followed by negative selection with antibodies can specifically be targeted to individual alleles and loci in turn.

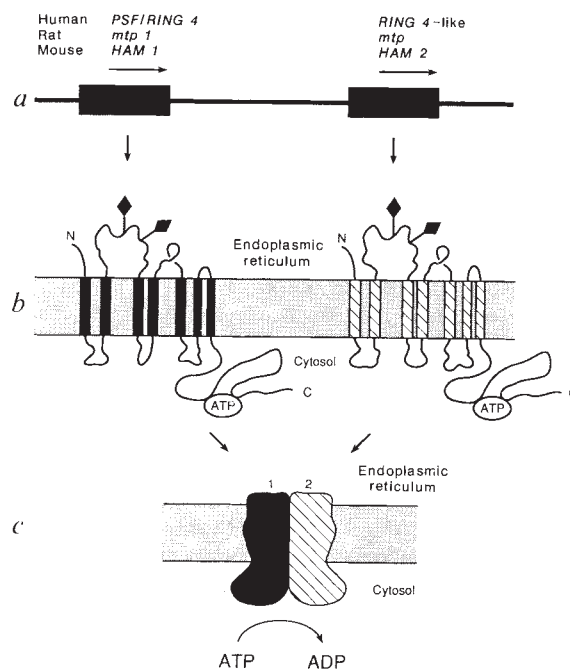
In this manner¹⁰ DeMars has created an incredible collection of MHC deletion and other mutants from the LCL721 B cell line. One group of mutants has a phenotype like that of the much-studied RMA-S and T2 cell lines — low steady-state levels of class I molecules at the plasma membrane which can be increased by addition of specific binding peptides^{11,12}. By analogy, Spies and DeMars attribute the defect to an insufficient ration of peptides in the endoplasmic reticulum, resulting from loss of a "peptide supply factor" (PSF), their sobriquet for the putative peptide transporter. Knowledge of the genes deleted in mutants with this phenotype previously led Spies *et al.*⁸ to identify the Y3 gene (identical to the human *RING 4* gene, and homologous to the mouse *HAM 1*, rat *mtp1* and the family of ATP-driven transporter genes — see figure) "as PSF with near certainty". This claim they now show is both half right, half wrong.

When Y3 — the PSF gene — was expressed in a mutant cell (LCL721.174), which has a large homozygous deletion encompassing Y3 and neighbouring genes, there was no increase in class I expression at the cell surface. But Spies and DeMars were able to transfect another recipient cell (LCL721.134) with less formidable mutations. In this cell one entire MHC haplotype is gone

but the only defect in the second haplotype is a non-functional Y3 gene. Expression of the normal Y3 gene in this mutant reversed its phenotype giving normal levels of class I molecules at the cell surface (of course only those from the second MHC haplotype). *Ergo* the product of Y3 is necessary but not sufficient for provision of peptide supply. Spies and DeMars now admit at least one other gene in the vicinity of Y3 is also involved. With hindsight this proposition seems rather probable, as the groups of Trowsdale⁷, Monaco⁵ and Howard⁶ documented the presence, in the MHCs of humans, mice and rats respectively, of a second transporter-like gene that is both homologous and closely linked to the first.

Members of the superfamily of ATP-driven transport proteins are found throughout phylogeny and in the course of evolution they have become specialized to pump a remarkable range of molecular structures across cellular membranes¹³. They have a common structure consisting of two ATP-binding domains and two hydrophobic transmembrane domains, which in different systems are contributed by between one and four polypeptides¹⁴. The two putative transporter genes found in the MHC encode polypeptides with one ATP-binding domain and one hydrophobic domain — half the usual functional unit. The pumps to be made from these genes are therefore likely to consist of either homodimers or heterodimers formed from their polypeptide products.

The results of Spies and DeMars can be



One model for the formation of a complete peptide pump from the proteins encoded by known transporter-like genes in the MHC. *a*, Two genes in the MHC of humans, rats and mice encode members of the superfamily of ATP-driven transmembrane transporters. *b*, Diagram of the membrane orientation of the MHC-encoded transporter based on Fig. 9 of ref. 15 and Fig. 4 of ref. 6. *c*, Diagram showing how the two gene products may form a heterodimer.

1. Spies, T. & DeMars, R. *Nature* **351**, 323-324 (1991).
2. Townsend, A. R. M., Gotch, F. M. & Davey, J. *Cell* **42**, 457-467 (1985).
3. Salter, R. D. & Cresswell, P. *EMBO J.* **5**, 943-949 (1986).
4. Powis, S. J., Howard, J. C. & Butcher, G. W. *J. exp. Med.* **173**, 913-921 (1991).
5. Monaco, J. J., Cho, S. & Attaya, M. *Science* **250**, 1723-1726 (1990).
6. Deverson, E. V. *et al.* *Nature* **348**, 738-740 (1990).
7. Trowsdale, J. *et al.* *Nature* **348**, 741-744 (1990).
8. Spies, T. *et al.* *Nature* **348**, 744-747 (1990).
9. Parham, P. *Nature* **348**, 674-675 (1990).
10. Kavathas, P., Bach, F. H. & DeMars, R. *Proc. natn. Acad. Sci. U.S.A.* **77**, 4251-4255 (1980).
11. Cerundolo, V. *et al.* *Nature* **345**, 449-452 (1990).
12. Hosken, N. A. & Bevan, M. J. *Science* **248**, 367-370 (1990).
13. Kuchler, K. & Thorne, J. *Curr. Opin. Cell Biol.* **2**, 617-624 (1990).
14. Hyde, S. C. *et al.* *Nature* **346**, 362-365 (1990).
15. Kuchler, K., Sterne, R. E. & Thorne, J. *EMBO J.* **8**, 3973-3984 (1989).

speculatively interpreted to mean that both putative transporter genes are essential for peptide supply to class I molecules and that "peptide supply factor" is in fact a heterodimer of the two gene products (see figure). Resolution of this question should help the search for a direct demonstration that the

putative transporter proteins do actually pump peptides. □

Peter Parham is in the Departments of Cell Biology, and Microbiology and Immunology, Stanford University School of Medicine, Stanford, California 94305, USA.

GAMMA-RAY BURSTERS

The answer is near

Charles D. Dermer

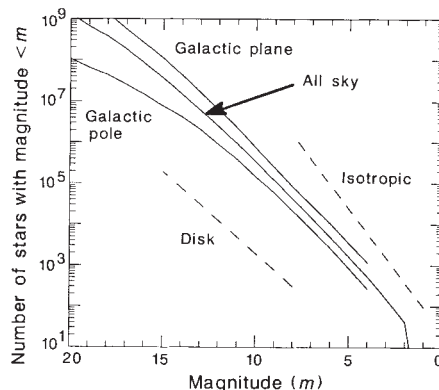
DEBATE has continued over the distances to the sources of brief, intense flashes of γ radiation called γ -ray bursts since their discovery was announced¹. The apparently isotropic distribution of burst directions on the sky has made it impossible to determine whether the bursts arise in our Galaxy or have an extragalactic or cosmological origin. Attempts to identify steady counterparts at other wavelengths have also been fruitless, except in the case of the unusual 5 March 1979 event which, if located in the Large Magellanic Cloud, would place 'normal' bursts far outside the Galaxy. On page 296 of this issue², a French-Soviet collaboration concludes, from an analysis of the statistical properties of burst data, that γ -ray bursters are nearby galactic phenomena. If this conclusion is substantiated, an exciting chapter in high-energy astrophysics will finally have come to an end.

Gamma-ray bursts are cosmic fireworks that flare up for a few seconds, emit the bulk of their energy in γ -ray photons, and then become invisible at all wavelengths, in most cases never to recur. Analyses of the spatial distribution of bursts are therefore necessarily statistical, involving the brightness time histories, positions on the sky and bursting rates. Because of the great variations in the brightness histories, these are usually replaced by a single value representing the burst strength, normally chosen to be peak flux or peak detector count rate.

The technique used by Atteia and colleagues², and some of the pitfalls such studies encounter, can be illustrated through the more familiar example of star counts in the night sky. The figure shows the visual magnitude distribution of all stars in the sky, as well as the magnitude distribution of stars in the direction of the galactic plane and pole³. (Magnitude is a logarithmic measure of intrinsic brightness, the faintest stars having the largest magnitude.) Slopes of isotropic and disk distributions of equal-brightness stars are shown for comparison. At magnitudes $m < 2$, the plots become limited by statistics: there are not enough stars brighter than second magnitude to conclude anything about their distribution. The all-sky distribution is roughly isotropic for $m < 5$, and corresponds to a disk distribution for $5 \leq m \leq 15$. At fainter magnitudes we are evidently seeing past the edge of the disk, where the star density becomes small.

The shape of the stellar distribution function depends on whether we are looking toward the galactic pole or plane. In the polar direction, we start to run out of stars with $m > 10$, whereas the edge of the disk distribution in the direction toward the galactic plane occurs for much fainter stars with $m > 15$. Given a representative source whose distance we know, such as the Sun, a reasonably accurate picture of the Galaxy as a flattened disk distribution, with vertical scale height of about 100 parsecs, can be constructed.

The approach by Atteia *et al.* is similar, but instead of magnitude, they use the variable $V/V_{\max}$ introduced by Schmidt⁴ in his studies



Magnitude distribution of stars observed in all directions (all sky), and in the direction of the galactic plane and pole. The density of stars toward the galactic plane and pole are scaled to the all-sky density and total number distribution.

of quasars. Here V stands for volume, and corrects the measured peak count rate for the increase, proportional to r^3 , in the number of sources at distance r and for the $1/r^2$ decrease in their apparent flux; $V_{\max}$ is the maximum volume that can be probed by a detector because of threshold limitations. The average of a sample of $V/V_{\max}$ values equals 0.5 if γ -ray bursts are uniformly distributed in space, and all detector biases are taken into account. In terms of the figure, the $V/V_{\max}$ test provides a measure of the deviation of a distribution of stars brighter than a limiting magnitude from one with slope 0.6, which is the slope of the isotropic distribution.

Normally the $V/V_{\max}$ test is made without reference to direction. The new approach in the French-Soviet study is to consider the directional properties of faint bursts only,

that is, those with large $V/V_{\max}$ values which presumably originate at large distances. By selecting only these faint distant bursts (or, referring again to the figure, only dim stars), the spatial anisotropies are magnified. The authors find statistically significant evidence for the dim distant bursts to lie in the galactic plane. Such a result, if true, goes a long way toward demolishing all extragalactic models of γ -ray bursts.

The approach seems sound: although the γ -ray bursts will have a broad range of luminosities, rather than the single 'standard candle' luminosity implicit in the analysis, it seems unlikely that the difference would alter the conclusions. And although obscuration by dust and gas can distort optical surveys, the effect would be negligible for γ -rays, and in any case would be most marked for bursts in the plane of the Galaxy. In focusing on the peak count rates from bursts, the authors of the study have chosen the measure of intensity shown^{5,6} to be most free from instrumental bias. The authors show that it is highly unlikely that the excess of bursts seen in the plane is a result of chance; although the analysis involves the low-intensity 'cut off' value of $V/V_{\max}$, one presumes that the conclusion does not depend on the particular dividing line between 'bright' and 'faint' bursts. Also, one might ask whether the enhancement in the galactic plane would have been seen if their approach had been applied to other catalogues of bursts, in particular the Konus catalogue with over 60 different localized bursts⁷⁻¹⁰.

Although there will always be sceptics in a subject as fraught with detector biases and statistical uncertainties as γ -ray astronomy, this work presents the strongest case to date that γ -ray bursters are galactic. Compelling evidence that the burst distribution is spatially limited has been known since the balloon test¹¹ of the prototype of the burst detector now flying on the Gamma Ray Observatory, so these results are by no means unexpected. The burst experiment on GRO is now recording γ -ray bursts at a frequency of about 1 per day (G. J. Fishman, personal communication), so confirmation or refutation of this study is not, perhaps like the bursters themselves, far away. □

Charles D. Dermer is in the Department of Space Physics and Astronomy, Rice University, Houston, Texas 77251-1892, USA.

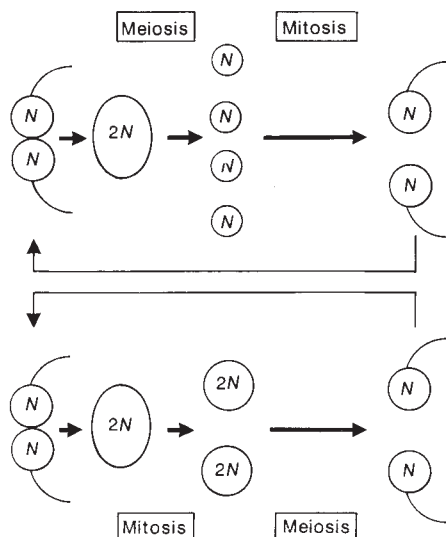
1. Klebesadel, R. W., Strong, I. B. & Olson, R. A. *Astrophys. J.* **182**, L85-L88 (1973).
2. Atteia, J. L. *et al.* *Nature* **351**, 296-298 (1991).
3. Allen, C. W. *Astrophysical Quantities* 3rd edn p. 244. (Athlone, London, 1976).
4. Schmidt, M. *Astrophys. J.* **151**, 393-409 (1968).
5. Mazets, E. P. in *19th Int. Cosmic Ray Conf.* (La Jolla) **9**, 415-430 (1986).
6. Paczyński, B. & Long, K. *Astrophys. J.* **333**, 694-699 (1988).
7. Mazets, E. *et al.* *Astrophys. Space Sci.* **80**, 3-83 (1981).
8. Mazets, E. *et al.* *Astrophys. Space Sci.* **80**, 85-117 (1981).
9. Atteia, J. *et al.* *Astrophys. J. Suppl.* **64**, 305-382 (1987).
10. Higdon, J. C. & Schmidt, M. *Astrophys. J.* **355**, 13-17 (1990).
11. Meegan, C. A., Fishman, G. J. & Wilson, R. B. *Astrophys. J.* **291**, 479-485 (1985).

When to be diploid

Brian Charlesworth

REPORTS in this issue^{1,2} by Kondrashov and Crow (page 314) and Perrot and colleagues (page 315) bring fresh insights to the long-standing question³⁻⁶ of the selective forces that favour the evolution of a prolonged diploid phase in the sexual life cycle. Both sets of authors explore the hypothesis that the advantage arises from the 'covering-up' of recessive deleterious mutations, which are fully expressed in haploids but only partially expressed in diploids^{5,6}.

Elucidation of the alternation of haploid



The postponement of meiosis (bottom) in a hypothetical unicellular species where initially meiosis immediately follows fusion of gametes (top). Vegetative cells are represented by circles, gametes by circles carrying curved lines, and zygotes by ovals. The level of ploidy of each cell is indicated by N or $2N$.

and diploid generations in plants was one of the triumphs of nineteenth-century cytology^{3,7}. In the most primitive unicellular algae, the diploid phase of the life-cycle is confined to a zygote that divides meiotically to produce haploid vegetative cells, which differentiate into gametes after a number of mitoses. A prolonged evolutionary process, in which the diploid phase has been extended at the expense of the haploid phase, has led to the system of reproduction of modern seed plants in which the haploid phase is a mere rudiment, which follows meiosis and immediately precedes gamete formation (see figure).

To understand the issues, consider a single locus with alleles A and a in a population with a dominant haploid vegetative phase. Let the relative fitnesses of these alleles in haploids be 1 and $1-s$. If the mutation rate per generation from A to a is u , the mean fitness of a haploid population is $1-qu \approx 1-u$ (refs 1,2,5). Imagine that a genotype arises which causes its carriers to defer meiosis,

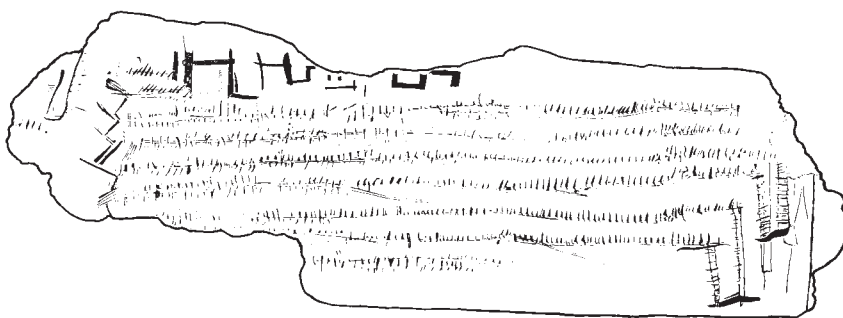
replacing the haploid phase with diploidy (see figure). Let the fitnesses of the diploid genotypes AA , Aa and aa be 1 , $1-hs$, and $1-s$. If h is non-zero, as suggested by the extensive data from *Drosophila*⁸, population fitness is chiefly affected by the elimination of heterozygotes (frequency $\approx 2q$), and the mean fitness of the new diploids is $1-2qhs \approx 1-2uh$. Provided that $h < 1/2$, so that the deleterious allele is at least partially recessive, the diploid fitness is greater than that of the haploids.

At first sight this provides a satisfactory explanation for the evolution of diploidy, because there is strong evidence both that $h < 1/2$ for deleterious alleles, and that a large number of loci are subject to mutation⁸. But this advantage may only be transient^{1,2,5}. If the diploid lineage fails to interbreed with the haploids, diploids will ultimately prevail over haploids only if their long-term equilibrium fitness is greater. But the equilibrium mean fitness of a diploid population is $1-2u$ (refs 1,2,5). Diploids end up being worse off than haploids, effectively paying the price of

carrying twice as many mutable loci.

The two new papers offer quite different solutions to this difficulty. Kondrashov and Crow¹ retain the assumption of non-interbreeding diploids and haploids, but assume that there are many loci with deleterious mutant alleles, and that individuals are eliminated by truncation selection (that is, anyone carrying a number of mutations exceeding a threshold value T is eliminated, whereas everyone else survives). In haploids, selection acts on the actual number of mutations carried by an individual, whereas in diploids incomplete dominance means that selection perceives only an h th of the number of mutations. It follows that the threshold number of mutations per individual for the diploid is approximately four times that for the corresponding haploids¹. The proportion of the population surviving truncation selection depends roughly on the distance between the population mean and the threshold. If the mean for haploids is n , then the proportions surviving are proportional to $T-n$ (haploids) and $T/h-4n$ (diploids). Thus diploidy confers a higher mean fitness if $h < 1/4$, a value that is close to that observed for deleterious alleles found in *Drosophila* populations⁸. This heuristic argument is confirmed by exact computer calculations, and the conclu-

Marking time in the Palaeolithic



A STRIKING claim, made by Alexander Marshack in the newly launched *Cambridge Archaeological Journal*, is that the enigmatic marks depicted above represent a non-arithmetic system of calendrical recording that dates to the Upper Palaeolithic of Europe some 10,000 years ago. The marks appear on the Taï plaque, of which the illustration reproduced here is a line drawing. The plaque is a fragment of rib 8.8 cm long which was discovered in a cave, in Drôme Department in Southwest France, in 1969.

Since its discovery, Marshack has doggedly examined the plaque in microscopic detail with a view to unravelling the meaning of the notations as a "cognitive form of visual problem-solving and structuring". His analysis leaned also on the examination and cross-comparison of other forms of pre-literate notation found in other cultures. Now, as a culmination of 20 years work — a point he is at pains to stress — Marshack argues that the plaque shows that the lunar year was understood, observationally if not arithmetically, by the end of the last Ice Age in Europe. It also carries evidence of the marking of solstices, he suggests, in which case there was at that time also an awareness of the solar year.

Marshack's article is innocent of statistical analysis, and he himself draws attention to the need for archaeoastronomers and ethnoarchaeologists to evaluate the data (no straightforward matter, as the long-running debate over the uses to which Stonehenge was put demonstrates). But should his interpretation find support, it will show that some of the impressive cultural features of Neolithic Europe draw upon indigenous tradition.

Tim Lincoln

sion also holds for somewhat less extreme forms of selection¹. Hence, if competition is sufficiently severe that selection is close to truncational, a long-term advantage to diploidy will exist.

Perrot *et al.*² take a quite different approach, arguing that diploids and haploids may interbreed by the fusion of haploid gametes which both types produce at the same time (see figure). Under this assumption, the frequency of the deleterious allele a will be approximately the same (say q') in both types. The diploids will have a fitness advantage if $h < 1/2$, because their fitness is $1 - 2q'sh$ instead of $1 - q's$. This conclusion is shown to be approximately correct by modelling a system of a locus subject to mutation and selection, together with a locus controlling the timing of meiosis.

Perrot *et al.* also make the interesting point that the mutational model of the advantage of diploidy assumes that deleterious mutations are partially recessive, even in a population recently derived from a haploid ancestor. For this to be true, Fisher's theory for the evolution of dominance of wild-type alleles by the modification of heterozygotes must be false, and dominance of wild-type must essentially be a developmental phenomenon. A survey of mutations of *Chlamydomonas reinhardtii* (H. A. Orr, personal communication) shows that the phenotypic effects of mutations are typically recessive, even in this classically haploid species. This is consistent with other evidence³.

Of course, these genetical arguments ignore ecological and physiological factors that may influence the relative advantages of diploidy and haploidy. For example, a haploid organism may have a growth advantage because of its smaller cell size¹⁰. The optimum length of the diploid phase of the life-cycle must reflect a balance between advantages of diploidy, and factors favouring haploidy. Given that the genetic advantage of diploidy is larger, the larger the per genome mutation rate¹, it is understandable that there is a broad correlation between organismal complexity and length of the diploid phase³, assuming that more complex organisms have larger mutable genomes. □

Brian Charlesworth is in the Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA.

The good companions

Lars Hernquist

MOST stars in our Galaxy are not isolated, as the Sun is, but are gravitationally bound to nearby companions¹. Were they born this way, or have companions been thrust upon them? In dense stellar environments, such as globular clusters, close stellar encounters can result in binaries. But in the Galaxy as a whole, stellar collisions are infrequent, and, in any event, they are not relevant to groups with more than two stars. On page 298 of this issue², Boss shows, with computational models, that aggregates of stars can be created during events similar to those responsible for forming individual stars. Furthermore, the structure of the resulting group depends mainly on the initial state of the gas from which the stars are born, so that studies of multiple-star systems may reveal properties of star-forming regions that would otherwise be inaccessible.

Much of the gas in galaxies like our own is contained in dense molecular clouds. It is generally believed that stars form from this material as it collapses in small regions that are gravitationally unstable. It has long been held that collapsing gas will fragment into a small number of bound subunits each of which eventually turns into a star³. But early numerical investigations of this fragmentation hypothesis were mostly inconclusive and never produced stable configurations with more than two members⁴.

All real, long-lived systems with more than two stars are organized hierarchically. For example, stable groups of three stars contain a tightly-bound pair and the third star orbits at a much larger distance. Multiple-star systems not configured in this way are dynamically unstable and eventually dissolve.

Protostars

In this issue, Boss uses a novel computational algorithm to show how hierarchical systems of protostellar objects may arise in the fragmentation picture. He examines the evolution of collapsing gas clouds in various initial states. In some cases, his model clouds evolve into thin bars which fragment into two independent entities which he identifies as protostellar cores. This effect, which has been seen in previous simulations⁵, may account for at least some galactic binaries. In other cases, rather than developing a bar, the clouds collapse into thin disks which subsequently fragment into two pairs of protostellar cores. Moreover, these cores are arranged hierarchically in the sense that they comprise two well-separated binaries, suggesting that a stable quartet of stars has been born.

Boss argues that the outcome of fragmentation depends sensitively on initial conditions in the gas. Clouds prone to yielding hierarchical systems with more than two stars may have been close to equilibrium and have

experienced a less violent evolution than those in which binaries form. If so, then statistical properties of multiple-star systems may provide us with an indirect probe of conditions in molecular clouds in the Galaxy.

Accretion

Although Boss's calculations are probably important, their ultimate significance will not become clear without further modelling. The masses of the protostellar cores in his quartets are 100–200 times smaller than those of stars like the Sun. For these cores to evolve into objects of solar mass, they must accrete more than 100 times their own mass from the gas which surrounds them. Moreover, even if the cores do eventually accrete these quantities of gas, their long-term survival as independent entities is not ensured. As Boss is careful to note, the interaction of the protostars with ambient gas may cause them to spiral towards one another and merge. The merger rates appropriate for conditions found in star-forming regions like those studied by Boss are unknown and can be estimated precisely only by further numerical simulation. Unfortunately, this task will be more demanding computationally than the feat that Boss has already accomplished: the timescales over which protostars merge are significantly longer than those in which the cores arise.

Perhaps most nettlesome is the issue of the detailed structure of the gas before its collapse. As in earlier studies, Boss assumes that the angular density profiles of clouds have a modest bi-symmetric component. It is certainly the case that the cores of molecular clouds are not smoothly distributed⁶. In general, one might expect the density to vary over a number of angular scales and it is not obvious that a bi-symmetric feature would be preferred. One possibility, examined most recently by de Felice and Sigalotti⁷, is that independent regions of collapsing gas might tidally force bi-symmetric responses in each other, just as paired ocean tides result on Earth from its interaction with the Moon. Indeed, many of de Felice and Sigalotti's calculations yield multiple-star systems though, apparently, none that are hierarchical. The lack of hierarchical systems may partly be due to the low resolution afforded by their numerical approach and it will be of some interest to investigate the evolution of tidally forced models using algorithms such as those

1. Kondrashov, A. S. & Crow, J. F. *Nature* **351**, 314–315 (1991).
2. Perrot, V., Richerd, S. & Valério, M. *Nature* **351**, 315–317 (1991).
3. Svedelius, N. *Bot. Gaz.* **83**, 362–384 (1929).
4. Stebbins, G. L. *Variation and Evolution in Plants* (Columbia Univ. Press, New York, 1950).
5. Crow, J. F. & Kimura, M. *Am. Nat.* **99**, 439–450 (1965).
6. Maynard Smith, J. *The Evolution of Sex* (Cambridge Univ. Press, 1978).
7. Farley, J. *Gametes and Spores* (Johns Hopkins Univ. Press, Baltimore, 1982).
8. Crow, J. F. & Simmons, M. J. in *The Genetics and Biology of Drosophila* (eds Ashburner, M., Carson, H. L. & Thompson, J. N.) Vol. 3c, 1–35 (Academic, London, 1983).
9. Kacser, H. & Burns, J. A. *Genetics* **97**, 639–666 (1981).
10. Cavalier-Smith, T. *The Evolution of Genome Size* (Wiley, Chichester, 1985).

1. Abt, H. A. *A. Rev. Astr. Astrophys.* **21**, 343–372 (1983).
2. Boss, A. P. *Nature* **351**, 298–300 (1991).
3. Hoyle, F. *Astrophys. J.* **118**, 513–528 (1953).
4. Shu, F. H., Adams, F. C. & Lizano, S. *A. Rev. Astr. Astrophys.* **25**, 23–81 (1987).
5. Boss, A. P. *Astrophys. J.* **319**, 149–161 (1987).
6. Myers, P. C. in *Interstellar Processes* (eds Hollenbach, D. J. & Trumpler, H. A. Jr) 71–86 (Kluwer, Dordrecht, 1987).
7. de Felice, F. & Sigalotti, L. D. G. *Mon. Not. R. Astr. Soc.* **249**, 248–261 (1991).
8. Fekel, F. C. Jr *Astrophys. J.* **246**, 879–898 (1981).

developed by Boss.

Despite these caveats, it should be emphasized that Boss's results provide the first direct evidence that fragmentation can produce hierarchical structures. Whether or not his protostellar cores remain distinct and evolve into actual stars will presumably be determined by future computations. For now, it is worth noting that quadruple-star systems are not uncommon. Their actual frequency is difficult to determine in practice owing to selection effects. Nevertheless, roughly 1 per cent of observed 'stars' may in fact be unresolved quartets¹, implying that

several per cent of all stars reside in groups of four. Many four-star systems with the hierarchical structure seen in Boss's models (distant pairs of tightly-bound binaries) have been detected visually⁸. To the extent that the current simulations continue to be relevant in the face of more detailed calculations, Boss's work may have already explained the formation of quadruple stars such as μ Ori, HR 3337 and ξ UMa. $\square$

Lars Hernquist is at the Lick Observatory, University of California, Santa Cruz, California 95064, USA.

OLFACTION

The strong scent of success

Doron Lancet

WITH their report in *Cell*¹, Linda Buck and Richard Axel may have at last laid a long-standing argument over the mechanisms of smell to rest. They describe the molecular cloning of a large family of proteins, each with seven transmembrane domains and all expressed specifically in olfactory epithelium. In all likelihood these are the long-awaited olfactory receptors.

The words 'long-standing' are appropriate. In the first century BC Lucretius suggested that 'atomic shapes' could underlie the sense of smell, and since then numerous theories of olfaction have been proposed. Some have argued that direct modulation of the lipid in membranes could mediate detection of odours²; on the other

hand, the existence of specific anosmias^{3,4} (the inability to detect particular odorants), and the discovery of an olfactory transduction cascade⁵⁻¹⁰, have made the case for stereospecific protein receptors a strong one. But most receptors are isolated by affinity-related techniques¹¹. In the olfactory system, agonist diversity and lipophilicity, and the lack of defined antagonists, have impeded progress⁷. Receptor identification by way of alternative criteria⁵⁻⁸ such as glycosylation, transmembrane disposition and tissue specificity, or through the isolation of olfactory mutants in *Drosophila*³, has been attempted without success.

Buck and Axel's remarkable discovery is an example of biological prediction come true. Olfaction is mediated by a GTP-binding-protein (G-protein) cascade⁵⁻⁹. It was therefore inferred^{6,8} that olfactory receptors probably belong to the G protein-coupled, seven-transmembrane-domain receptor superfamily. With this notion in mind, Buck and Axel carried out a first screen for rat olfactory receptor clones. They applied the polymerase chain reaction (PCR) to 30 different oligonucleotide pairs, derived from conserved sequences in transmembrane domains 2 and 7. This is an elaborate version of a technique previously used to fish for new G protein-coupled receptors in other systems¹². But as the process often yields a 'zoo' of novel as well as known receptors, a second selection rule had to be formulated.

Among the millions of possible odorants, many are evolutionarily 'unknown' to the organism until they are encountered for the first time.

Olfaction: molecules to network

Olfactory epithelium contains between 10 million and a 100 million sensory cells, and clearly all of them are not equal: they respond to different (though broad and overlapping) ensembles of smells (odorants). With the report by Buck and Axel it becomes possible to uncover the molecular basis for such individuality — most probably, each cell type expresses a subset of the receptor repertoire. Double labelling the sensory tissue with different receptor subtype-specific antibodies or oligonucleotides will help define how many different receptor genes are transcribed in each cell. If the answer is one (that is, clonal exclusion is at work, as in the immune system) then the known broad spectra of individual olfactory cells are wholly attributable to receptor protein 'promiscuity', a model testable by studies of functional expression.

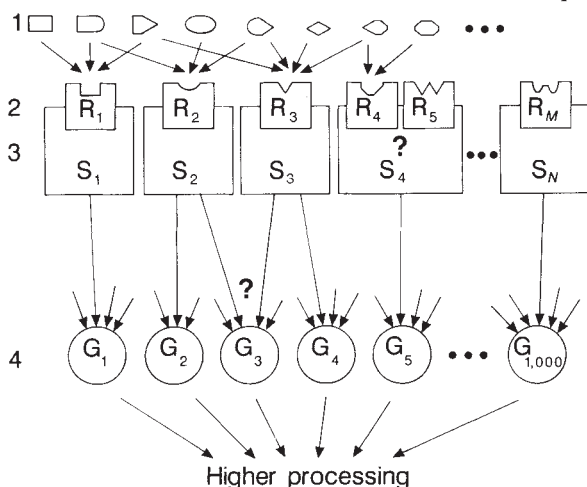
The crudest of the 'labelled lines' premises — that one cell type codes one odour — is largely out of fashion; rather, odour quality is thought to be coded by an 'across fibre' pattern of sensory neuronal firing^{14,15}. It has not been easy to discern this pattern experimentally, presumably because the different sensory cell types form a fine mosaic, with only minor segregation by odour specificity across the epithelial sheet. High-resolution labelling with receptor subtype-specific reagents could reveal the nature and grain of the presumed clonal distribution underlying the expected patterns of function in the epithelium.

A different situation prevails in the olfactory bulb, the first relay in the central nervous system, where individual glomeruli (100- μ m sized clusters of synapses between primary and secondary neurons) show a clear spatial pattern in their responses to a given odorant^{14,15}. Thus, axons from a given sensory cell type may project preferentially to one or a few glomeruli. Such a hardwired projection scheme, possibly rather invariant with time and among individuals¹⁴, embodies a more subtle version of the 'labelled lines' model. The entire pattern across units manifesting such discrete activity is then thought to code an odour.

To ask specific questions about projections to the olfactory bulb, it is necessary to identify unequivocally functional subgroups of sensory neurons. Responses to odours are too broad to be useful, and location in the epithelium is an uncertain guide to function. Now, one might achieve such definition by labelling sensory cells for their unique receptor(s), then find how selective their glomerular projection is. A tempting thought⁵ is that each glomerulus is a unique 'address' for a clone of sensory cells expressing a given olfactory receptor type. The number of glomeruli (about a thousand) is not unlike the extrapolated receptor count.

The synaptic target of a given sensory cell seems to depend on the receptor(s) they express. With the receptors cloned, it will be possible to determine the order of events in development and regeneration: do committed sensory cells recognize their target according to their predetermined receptor specificity, or is it contact with a synaptic target that specifies differentiation and commitment?

D.L.



Schematic view of olfactory neurons in the epithelium and their projection to olfactory bulb. (1) Estimates for the number of odorants range from 10,000 upwards. (2) Receptor proteins of M number of types are carried on (3) N types of sensory neuron; M is currently estimated as 100–1,000, and if clonal exclusion holds then $N = M$. Between stages (3) and (4) — (4) being the 1,000 or so glomeruli of the olfactory bulb — a convergence of some 100,000 sensory cells onto 1 glomerulus has to occur¹⁴. Open questions (?) are whether each neuron expresses more than one receptor type, and whether there is a one-to-one relationship between neuron types and glomeruli. The dual projection of S_2 and S_3 represents output from different individual neurons expressing the same receptor, as each sensory cell is known to synapse in only one glomerulus (cf. ref. 14).

An earlier insight was that olfaction, like immunity, could meet this challenge of xenobiotic recognition by randomly evolving a diverse repertoire of receptors (*cf.* ref. 5). But just how large the olfactory repertoire might be was a matter of conjecture — estimates ranged from several dozen⁴ (the number of specific anosmias known in humans) to several thousands⁵.

The crucial determinant in Buck and Axel's success was an insightful use of such anticipated diversity as a second selection criterion. They analysed each DNA fragment produced by PCR with frequently cutting restriction enzymes, and reasoned that any band concealing a multiplicity of forms would manifest a restriction complexity that was higher than expected. Rewardingly, only one of the many dozen PCR products examined displayed a high degree of such complexity. Out of 36 resultant library-selected clones, 18 were different, showing that a large family had indeed been uncovered.

Genomic Southern blot analysis provided a lower limit of 100 for the total number of related genes, and the authors forecast that the actual number may be several times larger, approaching the higher earlier estimates. Odour recognition could nominally be achieved by many fewer receptors^{5,8}, but redundancy, as elsewhere in the nervous system, might allow more elaborate processing in the central nervous system and better signal discrimination.

The candidate olfactory receptor family members embody a new species of seven-transmembrane-domain protein; although they have practically all the hallmarks of the superfamily¹¹, they express some new sequence motifs. Most distinctly, they are rather minimal in structure, having extremely short extracellular and cytoplasmic domains — could they be ancestral superfamily 'prototypes'? A future quest will be to understand how G-protein activation is nevertheless accomplished.

Further evidence that the receptors do indeed mediate olfaction is that the genes concerned are expressed specifically in olfactory epithelium, and their products are enriched in a preparation of putative olfactory neurons. Also of interest is the pattern of sequence diversity they manifest. First, there are several subfamilies, hypothetically related to ligand groupings. Second, transmembrane domains 3, 4 and 5 are more highly divergent than other regions of the protein. A diversified odorant-binding site could thus reside in the generic 'transmembrane barrel', which harbours contact residues for adrenergic and muscarinic ligands, as well as for retinal in rhodopsin¹¹.

What remains to be done to establish fully that these are the receptors for olfaction? First on the list is functional expression; that is, the demonstration that the polypeptide products can mediate odorant activation of the G-protein cascade. Caution should be exercised here, however, because little is known about odorant agonist and antagonist

'pharmacology' in most species. Extensive screening may be required to find the right one for a given receptor clone; and then, would merely showing activation by an agonist discovered by chance constitute proof? Species such as insects^{3,9} or aquatic animals¹⁰, in which physiological odorants (including pheromones) are better defined, and in which the receptor repertoire is perhaps smaller, may be the best with which to start.

Additional goals will be to correlate a behavioural phenotype, such as a specific anosmia, with a particular receptor genotype, as was elegantly accomplished for colour vision¹³; to confirm that the receptors are localized in the sensory neurons and that they segregate to their specialized ciliary extensions; and to find out the size and genealogy of this new multigene family of receptors.

And what of working out the underlying genomic organization? Like most other genes for receptors consisting of seven transmembrane domains, they seem to have no introns¹ — this will be a great help. The limited data suggest no obvious immunoglobulin-style gene rearrangements (or RNA splicing) as means of generating diversity within the polypeptide reading frame. But this does not preclude the possibility that clonally restricted expression of the receptors (see box on the previous page) proceeds through gene conversion to the vicinity of an olfactory-specific active promoter.

The cloning of genes for olfactory receptors now provides access to the molecular cell biology of olfactory coding, network formation and plasticity (see box and figure). This will be a key to understanding how a complex array of continuously regenerating sensory detectors conveys an unfailing 'image' of an odour to the brain^{14,15}. The sense organ that first gave mankind a glimpse of biological specificity may thus turn into a useful model system for investigations of neuronal wiring and information processing. □

Doron Lancet is in the Department of Membrane Research and Biophysics, The Weizmann Institute of Science, 76100 Rehovot, Israel.

1. Buck, L. & Axel, R. *Cell* **65**, 175-187 (1991).
2. Nomura, T. & Kurihara, K. *Biochemistry* **26**, 6141-6145 (1987).
3. Helfand, S. L. & Carlson, J. R. *Proc. natn. Acad. Sci. U.S.A.* **86**, 2908-2912 (1989).
4. Amore, J. E. *Ann. N. Y. Acad. Sci.* **237**, 137-143 (1974).
5. Lancet, D. A. *Rev. Neurosci.* **9**, 329-355 (1986).
6. Lancet, D. & Pace, U. *Trends biochem. Sci.* **12**, 63-66 (1987).
7. Snyder, S. H., Sklar, P. B., Hwang, P. M. & Pevsner, J. *Trends Neurosci.* **12**, 35-38 (1989).
8. Reed, R. R. *Cell* **60**, 1-2 (1990).
9. Breer, H., Boekhoff, I. & Tareilus, E. *Nature* **345**, 65-68 (1990).
10. Restrepo, D., Miyamoto, T., Bryant, B. P. & Teeter, J. H. *Science* **249**, 1166-1168 (1990).
11. O'Dowd, B. F., Lefkowitz, R. J. & Caron, M. G. A. *Rev. Neurosci.* **12**, 67-83 (1989).
12. Parmentier, M. *et al. Science* **246**, 1620-1622 (1989).
13. Nathans, J. A. *Rev. Neurosci.* **10**, 163-194 (1987).
14. Shepherd, G. M. in *Taste, Olfaction, and the Central Nervous System* (ed. Pfaff, D. W.) 307-321 (Rockefeller Univ. Press, New York, 1985).
15. Kauer, J. S. *Trends Neurosci.* **14**, 79-85 (1991).

Sperm warfare

MALES can compete for females in two distinct ways. In many birds and mammals, for example, the males compete either by direct aggression or by the acquisition and defence of territory. Some insects and vertebrates also go in for sperm competition. Each female mates with many males in rapid succession, and the struggle for descendants occurs between their sperm. The most active, vigorous and long-lasting sperm are most likely to fertilize her ova.

Daedalus now points out that even the most vigorous sperm are not perfectly competitive: any more than battleships would be with wonderful engines but no guns or armour. Such a decisive struggle, so close to the central pivot of reproductive success, must have evolved guns and armour as well. At first Daedalus imagined rival gangs of sperm slugging it out in the female reproductive tract like so many football supporters. But such small entities cannot really fight directly. Instead, sperm warfare must be chemical. Daedalus expects the semen of these species to be loaded with toxins or antigens to cripple rival sperm, while the sperm themselves will carry antibodies or other defences against return fire. The battle for survival and reproduction will not be so obvious, but will be just as deadly.

Now a recent social study suggests that sperm competition may have a place in human mating too. So DREADCO biochemists are looking for sperm competition in human males. They are mixing samples of donated semen under the microscope, and seeing which sperm survive. They hope to rank the donors in terms of aggressive fire-power, so to speak, and to identify their chemical weapons by sensitive immunoassay techniques.

Out of a large number of donors, one should emerge as 'top gun'. His sperm will kill or disable those of any other man. Daedalus expects him to be a repulsive, dispiriting, intensely dreary fellow, whose genes have specialized in this way for competing against more attractive males. The DREADCO team hope to identify the specific toxin deployed by his sperm, and synthesize it in quantity by chemical or genetic-engineering methods. DREADCO will then launch it as a new and powerful spermicidal contraceptive.

It should transform the market. As an entirely natural product it will evade papal denunciation. As the outcome of millions of years of ruthless evolutionary sexual competition, it will be effective, safe and free of side-effects — a potential father who harmed his mate would be severely handicapped in the race for genetic survival. Its only disadvantage is that it may not give perfect protection against the paternal efforts of certain rather boring and unattractive sexual partners. David Jones

Degradation of C₆₀ by light

SIR — Using production, extraction and purification procedures published previously^{1,2}, we have prepared numerous batches of pure C₆₀. Freshly prepared samples dissolve readily in benzene to give the deep magenta solutions that we have described², but samples stored without exclusion of light or oxygen will not redissolve completely — a reddish-brown deposit remains. Moreover, benzene solutions of C₆₀ exposed to light slowly produce a small amount of brown deposit. Digestion with benzene of solid samples of pure C₆₀ that have been stored in this way leads to successively paler extracts with an increasingly pinkish hue, we believe owing to the presence of traces of the decomposition product.

We have also observed that during column chromatography of mixtures of C₆₀ and C₇₀ on neutral alumina using hexane as eluent, the magenta colour from C₆₀ on the column is gradually lost if elution is interrupted and the column exposed to light. Both observations suggest the occurrence of some kind of photochemical reaction. We have confirmed this by irradiation of solutions of C₆₀ in HPLC-grade hexane with a water-cooled medium-pressure silica-jacketed Hanovia insertion ultraviolet lamp. We monitored the disappearance of C₆₀ by ultraviolet spectroscopy; reaction is complete within 10–16 hours. A buff precipitate is formed which consists of a mixture of products possessing the following features.

(1) Mass spectroscopy indicates that the fullerene cage is substantially or completely fragmented, depending on the reaction conditions. Fragmentation is less if the reaction is carried out under nitrogen (whence the oxygen content of the solution is reduced but remains significant). The products are not polymeric.

(2) The unprocessed product typically has a hydrogen content of 3–4 per cent and an oxygen content (determined by difference) of 40–54 per cent, depending on conditions; nitrogen is absent. It dissolves readily in polar solvents.

(3) Infrared and NMR spectra indicate that the following groups are present: C–H (including aromatic C–H), C=O (very intense), C–O, C=C, substituted aromatic or alkene, C≡C, C₂H₅, CH₂, alkyl–O, and CH₂–O (substantial). The product is partly soluble in potassium bicarbonate, indicating the presence of a carboxyl group.

Confirmation that hydrogen is abstracted from the solvent was provided by irradiation of a solution of C₆₀ in benzene, reaction being slower. Infrared spectra indicated the absence of OH groups, but C–H (including aromatic) and C=O bands were found.

Irradiation of a C₆₀ solution in hexane with a glass-enveloped medium-pressure lamp resulted in only trivial loss of colour after 16 hours, confirming that photochemical degradation requires ultraviolet irradiation.

Five conclusions follow from our work. First, aldehydes and/or ketones are formed, implying reaction with ozone. Given the accepted structure of the intermediate ozonide³, oxygen may end up within the cage, before cleavage to give the product.

Second, C₆₀ may be ultraviolet-sensitive in the absence of oxygen, and may thus partly decompose in the reaction vessel in which it is prepared (see, for example, refs 4,5) owing to the intense ultraviolet irradiation emanating from the carbon arc. Substantially higher yields of C₆₀ may therefore be achievable by using a flow system.

Third, samples of C₆₀ should be stored in the dark and under vacuum or nitrogen. Solutions of C₆₀, especially in either alkanes or cycloalkanes, should also be kept in the dark.

Fourth, although the alkyl–oxygen bonds present may in part arise from epoxides, NMR results show that three-membered carbon chains at least are also involved, and the most probable source of these is the sol-

vent. Use of photochemical irradiation and suitable co-reagents may therefore lead to a whole range of derivatives of C₆₀.

Finally, the reason for the failure hitherto to observe naturally occurring C₆₀ on Earth now becomes clear. (We have also failed to find it in chimney soot.)

ROGER TAYLOR

JONATHAN P. PARSONS

ANTHONY G. AVENT

STEVEN P. RANNARD

T. JOHN DENNIS

JONATHAN P. HARE

HAROLD W. KROTO

DAVID R. M. WALTON

*School of Chemistry and Molecular Sciences,
University of Sussex,
Brighton BN1 9QJ, UK*

1. Krätschmer, W., Lamb, L. D., Fostiropoulos, K. & Huffman, D. R. *Nature* **347**, 354–358 (1990).

2. Taylor, R., Hare, J. P., Abdul-Sada, A. K. & Kroto, H. W., *J. Chem. Soc. Chem. Commun.* 1423–1425 (1990).

3. March, J., *Advanced Organic Chemistry* 3rd Edn 1066–1072 (McGraw-Hill, London 1985).

4. Hare, J. P., Kroto, H. W. & Taylor, R. *Chem. Phys. Lett.* **177**, 394–398 (1991).

5. Haufler, R. E. *et al. Mat. Res. Soc. Proc.* (in the press).

Discrepancies in AIDS virus data

SIR — In early 1983 we described a new human retrovirus, then called LAV, now called HIV, associated with lymphadenopathy in a patient at risk of AIDS¹. We tried to propagate this isolate in continuous cell lines such as CEM and B-lymphoblastoid cell-like RAJI without any success, as reported at the first European conference on AIDS in Naples on 25 June 1983 and subsequently published². Clearly, this virus would grow only in peripheral blood lymphocytes (PBL), in cord blood lymphocytes and in bone marrow mononucleated cells.

We propagated this isolate exclusively in PBL for more than a year, and in 1984, we

adapted this virus to Epstein–Barr virus-transformed B-cell line³. The virus was subsequently cloned⁴ and sequenced⁵. Nothing was known about the genomic variability of various HIV isolates in 1983.

We found several discrepancies concerning the origin and the growth of this virus either in PBL or in continuous cell lines. Molecular analysis of sequence from previous isolates showed differences between the cloned and sequenced HIV⁶. We suggested that at least two variants could be present in the PBL-passed BRU and decided to analyse the samples biologically as well as molecularly. We used the HIV-BRU virus (JBB-LAV, January 1984) continuously propagated in PBL from an HIV-1-negative donor and took the sample maintained from June 1983 to January 1984 in these cells.

BIOLOGICAL PROPERTIES OF ISOLATES DERIVED AFTER GROWTH IN INDICATED CELLS

	BRU PBL 84 (JBB LAV)	Early passage of JBB-LAV in CEM	LAV 90 continuous passage in CEM	Molecular clone of LAV (transfected)
OKT4 A in PBL (5 µg ml ⁻¹)	No	No growth	Yes	Yes
OKT4 A in CEM (5 µg ml ⁻¹)	Yes	Yes	Yes	Yes
Soluble CD4 in PBL (5 µg ml ⁻¹)	No	+/-	Yes	Yes
Size of syncytia in MT4	Small	Normal	Large	Small
Neutralizing anti- bodies to HIV1-LAV	No	Yes	Yes	Yes
V3 loop (RIPA)	Absent	Present	Present	Present
TCID ₅₀ in PBL	10 ⁻⁵	< 10 ⁻¹ No growth	10 ⁻¹	
CEM	< 10 ⁻¹	10 ⁻²	10 ⁻³	
MT4	10 ⁻⁵	10 ⁻¹	10 ⁻³	

Virus stocks, titration pattern, block of the entry with OKT4 A or soluble CD4 are described in ref. 10; neutralizing-antibody determination in ref. 11. OKT4 A (Ortho Diagnostics), soluble CD4 was obtained from ANRS (D. Klatzmann). Monoclonal antibodies against V3 were from ANRS hybridolab Institut Pasteur (J. C. Mazie). No, no block of virus growth by OKT4 A or soluble CD4 neither neutralization by antibodies. Yes, block and neutralization.

We established this virus to continuous CEM cell line and compared its biological properties with the original isolate maintained in PBL. We tried to isolate each type of virus using OKT4 A or soluble CD4 for selection. The results, summarized in the table, suggest that one virus can be selected by the cultivation in continuous cell lines (CEM, H9, HUT 78 or B-lymphoblastoid cell), and the second in the PBL culture by selection using soluble CD4 or OKT4 A⁷. Molecular analysis of cloned and sequenced PCR product of the V4 and adjacent domains of BRU 84 PBL (JBB-LAV) revealed at least two types of clone, one corresponding to the published sequence by Wain-Hobson *et al.*⁵ and another to deletions of V4 domain characteristic of the sequence of Guo *et al.*⁶. After passage in CEM only one type of clone, corresponding to the sequence described in ref. 5, was found.

These results indicate that in the JBB-LAV preparation two viruses at least were present. We do not know whether they were present from the beginning in the same patient, as was described in other cases^{8,9} and selected by cultivation in PBL amplifying the

second virus (not detected in June 1983 at the first attempt to grow this virus in CEM²), or whether there could have been contamination with another isolate made in our laboratory which was not equipped with P3 facilities.

JEAN-CLAUDE CHERMANN
GHISLAINE DONKER
NOUARA YAH
DANIELLE SALAUN
NADIA GUETTARI
ODILE GAYET
IVAN HIRSH

INSERM U 322,
Campus Universitaire de Luminy,
BP 33, 13273 Marseille Cedex 9,
France

1. Barre-Sinoussi, F. *et al. Science* **220**, 868-871 (1983).
2. Chermann, J. C. *et al. Antibiot. Chemother.* **32**, 48-53 (1984).
3. Montagnier, L. *et al. Science* **225**, 69-72 (1984).
4. Alizon, M. *et al. Nature* **312**, 757-760 (1984).
5. Wain-Hobson, S. *et al. Cell* **40**, 9-18 (1985).
6. Guo, H.G. *et al. Nature* **349**, 745-746 (1991).
7. Rey, F. *et al. AIDS Forsch.* (in the press).
8. Von Briesen, H. *et al. J. med. Virol.* **23**, 51-66 (1987).
9. Fischer, A. G. *et al. Nature* **344**, 444-447 (1988).
10. Rey, F., Barre-Sinoussi, F., Schmidtmayerova, H. & Chermann, J. C. *J. virol. Meth.* **16**, 239-249 (1987).
11. Rey, F., Donker, G., Hirsch, I. & Chermann, J. C. *Virology* **181**, 165-171 (1991).

Hope for photonic bandgaps

SIR — The search for photonic bandgaps has not bitten the dust, John Maddox suggests¹. We have recently created three-dimensionally periodic dielectric structures which are to photon waves as semiconductor crystals are to electron waves. That is, these photonic crystals have a photonic bandgap, a band of frequencies in which electromagnetic waves are forbidden^{2,3}, irrespective of their direction of propagation. Photonic bandgaps inhibit the spontaneous emission of the frequencies concerned and allow for a new class of electromagnetic microcavities.

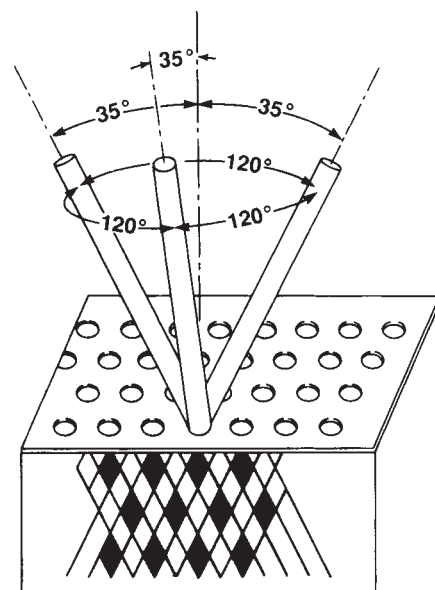
The creation of a photonic bandgap has proved to be more difficult than expected. It was clear from the outset that a face-centered-cubic (f.c.c.) array in real space would produce the roundest brillouin zone in reciprocal space. But what should be the shape of the atoms? Early suggestions were for cubic atoms², then spherical atoms and spherical voids⁴. There has been a long search for that optimal three-dimensional dielectric geometry favoured by nature and by Maxwell's equations. Most of this work has been done with macroscopic dielectric structures, when the forbidden wavelengths are in the microwave region.

During this same period, electronic band theorists began calculating photonic band structure. It rapidly became clear that the familiar scalar-wave band theory, so frequently used for electrons in solids, was in utter disagreement with experiment on photons. Recently, a full vector-wave band theory^{5,6} became available, which not only agreed with experiment but highlighted some discrepancies in experiment. Vector-

wave band theory showed that the spherical void structure allowed a degeneracy between valence and conduction bands at the W-point of the brillouin zone, resulting in a pseudogap, rather than a full photonic bandgap. The solution to this symmetry-induced degeneracy problem is to make the atoms nonspherical.

The figure shows a practical, new, f.c.c. structure, which simultaneously solves the two outstanding problems in the fabrication of photonic band structure. In this new geometry, the atoms are not spherical, but, rather, are distorted along $\langle 111 \rangle$, lifting the degeneracy at the W-point of the brillouin zone and permitting a full photonic bandgap rather than a pseudogap. Furthermore, this fully three-dimensional f.c.c. structure lends itself readily to microfabrication on the scale of optical wavelengths. It is created by simply drilling three sets of holes 35.26° off vertical into the top surface of a solid slab or wafer, as can be done for example by reactive ion-etching. At refractive index $n \sim 3.6$, which is typical of semiconductors, the three-dimensional forbidden photonic bandgap width, calculated and measured, is about 20 per cent of its centre frequency. Calculations indicate that the gap remains open for refractive indices $n \geq 2$.

If the $\langle 111 \rangle$ distortion is severe enough, the result can be the diamond structure⁷, which appears to give the widest photonic bandgaps of all. But diamond structure is difficult to microfabricate. It requires three



The method of constructing an f.c.c. lattice of nonspherical atoms. A slab of material is covered by a mask consisting of a triangular array of holes. Each hole is drilled through three times, at an angle 35.26° away from normal, and spread out 120° on the azimuth. The resulting criss-cross of holes below the surface of the slab, suggested by the cross-hatching shown here, produces a fully three-dimensionally periodic f.c.c. structure. The drilling can be done by a real drill bit for microwave work, or by reactive ion-etching to create an f.c.c. structure at optical wavelengths.

additional drilling directions in addition to those shown in the figure, all in the plane of the slab.

E. YABLONOVITCH

Bell Communications Research,
Navesink Research Center,
Red Bank, New Jersey 07701-7040,
USA

K. M. LEUNG

Department of Physics,
Polytechnic University,
Brooklyn, New York 11201, USA

Scientific Correspondence

SCIENTIFIC Correspondence is a relatively informal section of *Nature* in which matters of general scientific interest, not necessarily those arising from papers appearing in *Nature*, are published. Because there is space to print only a small proportion of the letters received, priority is usually given according to general interest and topicality, to contributions of fewer than 500 words and five citations, and to contributions using simple language.

If new results are being described, priority is generally given to communications that do not describe work in which the author is involved. Authors of contributions of this type should explain in a covering letter why theirs has a particular claim on *Nature's* space. Contributions may be sent to referees and, in the case of matters arising from material published in *Nature*, are sent to the author of that article for comment. A more detailed guide to authors is available from Washington or London. □

1. Maddox, J. *Nature* **348**, 481 (1990).
2. Yablonovitch, E. *Phys. Rev. Lett.* **58**, 2059 (1987).
3. John, S. *Phys. Rev. Lett.* **58**, 2486 (1987).
4. Yablonovitch, E. & Gmitter, T. *J. Phys. Rev. Lett.* **63**, 1950 (1989).
5. Leung, K. M. & Liu, Y. F. *Phys. Rev. Lett.* **65**, 2646 (1990).
6. Zhang, Z. & Satpathy, S. *Phys. Rev. Lett.* **65**, 2650 (1990).
7. Ho, K. M., Chan, C. T. & Soukoulis, C. M. *Phys. Rev. Lett.* **65**, 3152 (1990).

More English than most

Alan Mackay

Selections and Reflections: The Legacy of Sir Lawrence Bragg. Edited by John M. Thomas and David Phillips. *Science Reviews*: 1991. Pp.308. £28.50, \$57.

WILLIAM Lawrence Bragg (1890–1971) was the most English of English scientists and, to mark the centenary of his birth, the Royal Institution, of which Bragg was himself once director (as was his father, W. H. Bragg, before him), has published *Selections and Reflections*, a collection of comments and reminiscence by some 25 distinguished contemporaries, introduced by David Phillips' masterly biographical memoir and accompanied by some of Bragg's most characteristic papers and writings. The role of the Royal Institution in British science is itself illuminated.

The contributions of ten Nobel laureates and a dozen other distinguished colleagues of Bragg offer sharper vignettes of themselves than of Bragg, but provide colour for the more sober biography.

Bragg received the Nobel prize in 1915 at the age of 25, jointly with his father. He heard the news when he was serving on the Western Front developing a sound-ranging system (while German physicists were doing the same on the other side).

Together, the Braggs had discovered the principles of X-ray diffraction analysis of the arrangements of atoms in crystals, which today lie at the basis of our understanding of mineralogy, chemistry and molecular biology. In this way we can now see atoms and molecules.

Bragg was fortunate in his personal circumstances. He was recognized as a gifted child; his father was able to place him at the research frontier; he survived the war (where his younger brother did not). The Bragg family became part of the curious intermarrying dynastic system of Huxleys and Darwins which dominated British science. But why was Bragg so successful in solving problems?

Aaron Klug reports the key question (put by German physicists to R. W. James in Leipzig in 1931), "Tell me, how does Bragg discover things? He doesn't know anything" — meaning that, unlike Peter Debye, for example, he could not immediately, without notes, write down the quantum-mechanical hamiltonian for the electromagnetic field. Is there an English style in physics? This collection enables us to attempt an answer.

André Guinier, looking from the continental tradition, says that in France the mathematical solution of a problem seems more rewarding than the realization of an apparatus that works. He admits that, "in spite of the ingenuity of our models and the success of our logic, in many cases the progress of physics requires, and will always require, an empirical approach."

Max Perutz also, better able than anyone to compare Bragg's style with the continental, attributes success to "Bragg's astute powers of penetrating through the apparent complexities of physical phenomena to their underlying simplicity".

There are set-piece battles in science, like the structure of haemoglobin (to which Perutz became committed with eventual



Bragg — the likelihood of a problem meeting its answer in his mind was high.

victory), fusion energy (still very uncertain), targets identified and invested, and there is guerilla warfare where, having limited resources one reconnoiters the no-man's land between 'problems looking for solution' and the 'solutions looking for problems' to see where a junction might possibly be effected. This strategy is enormously more productive if the aim is to solve something (rather than to solve one particular problem).

Bragg had an excellent three-dimensional, visual-physical understanding of how things worked — not just a few things, but thousands of things. His science was the culmination of the Victorian ethos, when Britain was the workshop of the world. For example, the Society for the Propagation of Christian Knowledge had even published a book, *A Catechism of the Steam Engine* (as well as a manual of crystallography). Bragg had a huge number of concrete models and analogies 'to hand' in his mind. These he could

deploy and manoeuvre with great appropriateness. He obviously had a similarly large fund of questions and problems and enjoyed looking at things, like traditional industrial processes, which were full of stimulating problems. His father had produced a book, *Old Trades and New Knowledge* (1925), which had been the Royal Institution Christmas lecture of 1925. It covered the trades of the sailor, smith, weaver, dyer, potter and miner. Earlier, W. H. Bragg had taken Lucretius' title *On the Nature of Things* for a similar book. Chance favours the prepared mind and Bragg's was well prepared. The likelihood of a problem meeting its answer in his mind was high.

Curiously, perhaps significantly, and in contrast to Einstein and many of the great German figures, Bragg is reported as having absolutely no ear for the abstract patterns of music (neither had J. D. Bernal). He had an unusually good 'mind's eye' but not a 'mind's ear'. Nevertheless, Colin Blake compared his Englishness to that of Elgar.

Together, the Braggs had the superb vision that the microworld was just like the everyday world, only smaller. With this idea, Bragg revolutionized chemistry, metallurgy, mineralogy and later biology. With his examination of crystal structures, the atomic level of matter rose into sharp focus. Silicate chemistry was given a sensible basis. Macroscopic properties could be understood in terms of the arrangement of atoms, just as Lucretius and Democritus had guessed. Atoms became real — not like the monads of German *Naturphilosophie*.

The core of Bragg's vision was generalized microscopy. In 1929 he had represented crystal structures as pictures synthesized by adding the Fourier components of the electron density. In 1939, he wrote of the 'X-ray microscope' for seeing atoms, and, with the concept of the two-wavelength microscope (due to Martin Buerger) in the air and in dialogue with Dennis Gabor, Bragg almost developed the hologram. 'Microscopy by reconstructed wavefronts', his 1950 letter to *Nature*, is reproduced in *Selections and Reflections*.

Today, the model of matter with mechanical atoms has run into difficulties. The microworld is strange and not like our mesoworld. Even Richard Feynman did not feel at home in it:

We have always had a great deal of difficulty understanding the world view that quantum mechanics represents. At least I do, because I'm an old enough man that I haven't got to the point that this stuff is obvious to me. Okay, I still get nervous with it . . . You know how it always is, every new idea, it takes a generation or two until it becomes obvious that there's no real problem. I cannot define the real problem, therefore I suspect there's no real problem, but I'm not sure there's no real problem.*

Bragg never got used to it but, like the

* Feynman, R. *Int. J. theor. Phys.* **21**, 471 (1982); quoted by N. D. Mermin in *Physics Today*, April 1985.

guerilla who survives, did not venture into exposed regions where he was not comfortable. He stayed with small science, with classical atoms and with linear optical phenomena and their mathematics, as gradually heavy theoretical and physical apparatus became predominant. Nevertheless, when he was Cavendish Professor he encouraged radio astronomy and did not hinder nuclear physics.

Bragg was indeed fortunate in that his abilities matched the challenges and the times, however, it was not accident but skill that was necessary in assessing the abilities and the challenges and pitting one against the other. We see him, in about 1956, like Napoleon in his tent, privately estimating the coordinates of the heavy atoms in myoglobin using his own methods (where J. Kendrew was using the earliest available computer) and concluding that victory was then certain in the campaign to find the positions of the atoms in the first protein.

A number of Bragg's most characteristic papers (some rather difficult to find elsewhere) are reprinted in *Selections* and we may enjoy reading his advice on how to give a lecture about science (from *Marine Technology*, 1967). From this we realize that the apparently easy performance of lectures was the result of much thought on presentation and style. Apparent simplicity is not just simple, but is the result of the paring away of inessentials and looking at the relationships between a few key components. Nevertheless, as in a computer program, these few components may form one level in a hierarchic structure of considerable overall complexity.

We may ask today: are there still big problems in science where simplicity will get you the answer? Bragg's strategy was "Do not follow the fashion and do not hesitate to try an experiment which the theorists hold to be stupid."

Probably there are such problems, because, for example, computing and mathematics have been simplified by computer packages and there are huge databases which may be reckoned as simple. But they are harder to find and we need insight to see them. Somehow this is what Bragg had in abundance and what, in his lectures, he managed to communicate to others. As an undergraduate I heard him explain the birefringence of calcite with the excitement of someone who had realized only yesterday, and not a decade before, how the carbonate group worked in this respect. It is worthwhile studying how all this came about. Bragg was indeed fortunate that one paradigm carried him right through for 50 years, from the discovery of X-ray diffraction to the solving of the first protein structures. But he also had the vision to see the way ahead at each step, avoiding intractable complexity. □

Alan Mackay is in the Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK.

Big science biology

Kevin Davies

Mapping the Code. By Joel Davis. Wiley: 1991. Pp.294. £14.95, \$21.95.

AROUND the time of the Human Genome II conference in San Diego last year, the science correspondent for one of the major US television networks introduced a short item for the daily breakfast-TV programme on the goals of the Human Genome Project. It started promisingly enough — the familiar scene of a modern laboratory setting, people in white coats looking quizzically at autoradiographs, steady hands loading gels with purple markers; in short a typical day in the laboratory. Some carefully selected sound bites revealed that humans have about 100,000 genes, and that the aim of the project is to map them and eventually elucidate their function, with unimaginable rewards in the battle against diseases such as Alzheimer's and cancer.

But as the reporter handed back to Joe Garagiola, the long-retired ex-baseball star, in the New York studio the debate floundered as the camera caught the studio lights shining off his head; it seemed that what Mr Garagiola was more interested in was whether there was a gene for baldness, and when the scientists might find it. Joe and the reporter laughed for a few moments over the satellite link-up before the station abruptly cut to a commercial. As in-depth documentaries go, it may not have been much, but for the American public at least, it was a start.

In more academic circles, of course, arguments about the scientific merits and funding of the genome project have raged for years. The original notion of sequencing megabases of 'junk' DNA has been scotched, as most scientists seem to agree that mapping the human genome first and then sequencing the interesting regions, is the most sensible way to sell the project and its vast cost (by biology standards at least) to the politicians, researchers, and the public at large.

Given the enormous scale and estimated cost of the project, it is no surprise that a number of books professing to reveal the true workings of the project have appeared — at least three in the last six months. The dust jacket of Joel Davis's *Mapping the Code* (which contains by far the most colourful prose in the book) asserts that "Davis manages the impossible: he mixes hard science with human drama and history," but his analysis of the science is far poorer than managed in *Genome*, by Jerry Bishop and Michael Waldholz (Simon and Schuster, 1990: for review see *Nature* 348, 205; 1990), and in contrast to Lois Wingerson's readable and entertaining *Mapping Our Genes* (Dalton, 1990), virtually all signs of human life are

sadly lacking. The only remotely revealing insight that Davis uncovers is Paul Berg's knack of winning several cases of wine from distinguished colleagues on bets ranging from the extent of alternative splicing to the significance of 'junk' DNA.

Mapping the Code tackles the evolution of the genome project in greater detail than its predecessors, but debates such as which agency should coordinate the project and who should lead it simply do not make for enthralling reading. Much of the early portion of the book, in which Davis sketches the molecular genetic methods involved, is culled from congressional reports. The middle section details the meetings in the mid-1980s that led to the proposal to map the genome, drawing on numerous articles published in *Science* magazine at the time. But in Davis' hands, any sense of excitement or controversy is buried by his zeal in highlighting every meeting that took place, together with long (and tedious) lists of the participants, not to mention their affiliations. When he does become animated, it is over wholly unimportant subjects, such as whether the genome project should be classified as 'small Big Science' or 'big Small Science'.

The last third of the book is an improvement, as Davis touches on some of the important questions arising from the project itself: who owns the genome and its constituent chromosomes? How and when will we develop the supercomputers needed to analyse the bounty of information that the project will produce? And what are the potential uses and misuses of our own individual genetic map, assuming (as Davis does) that in 50 years or so we might be carrying around floppy disks with this very information?

For the most part, however, Davis shies away from a discussion of how the project should be structured. Is a 'top-down' format the most effective way of reaching our goal of a complete genetic map, or is it best to encourage the innovative investigators, as some such as Don Brown continue to argue? In the year or so since *Mapping the Code* was finished, genes responsible for some forms of retinitis pigmentosa, Alzheimer's disease and colon cancer have been identified, to name but a few. It is especially ironic that a fundamental understanding of Alzheimer's disease has long been promoted (especially by Jim Watson) as one of the reasons to map and sequence chromosome 21, and yet one mutation already appears to have been identified, and other gene loci are being located.

No matter how the map eventually comes about, the important task is that the genome project should help in the identification of genes that either cause or predispose one towards cancer, mental illness, and — of extreme importance to some — even baldness. □

Kevin Davies is an Assistant Editor of *Nature*.

High lights

Mark Saunders

Physics of Space Plasmas: An Introduction. By George K. Parks. Addison-Wesley: 1991. Pp.560. £40.45, \$48.50.

FEW natural spectacles can match the transient glory of the Sun's corona at solar eclipse, the sight of a bright comet plasma tail or the shimmering lights in a colourful auroral display. These vivid examples illustrate just some of the countless plasma phenomena within the Solar System that are included in the new field of space plasma physics. Fundamental to the subject's existence is the solar wind, a tenuous, fast-flowing, highly electrically conducting plasma blowing continuously out from the Sun. The solar wind gusts through the whole Solar System; its subtle interplay with planetary magnetic fields and planetary plasmas forms the basis for space plasma physics. Despite the need for further exploration, with exciting prospects ahead, space plasma physics has, after more than 30 years, now reached a mature and sophisticated level. With this maturity the time is ripe for the subject to become available as part of a science undergraduate's basic training.

Physics of Space Plasmas: An Introduction by George K. Parks is the most worthwhile attempt yet to provide a textbook aimed at this audience, and as such it provides a timely and welcome addition to the literature. The textbook is attractively produced, adequately illustrated and is written in a fairly readable style, the only marring of this good standard of production being an unfortunate page numbering error in the contents section after page 307. The book's primary focus is the theoretical concepts and techniques used for describing plasma processes in space, the latter including currents, boundaries, waves, shocks and instabilities. In each case, reference is given to experimental observations, although these form only a minor part of the book's scope. As befits a theoretical physics textbook, mathematics is widely used, though at times its level seems more appropriate to advanced post-graduates or to research workers. Indeed, considering the book's perhaps excessive length it may have been advisable to omit some of the more specialized theory and mathematics. Problems given after most chapters form an important and useful feature of the book.

Criticizing a book is never easy, especially

when the overall standard is good, and when different readers will doubtless have their own favourite topics that they wish to see emphasized. But in my opinion the textbook suffers, as is often the case, by reflecting the special interests of the author instead of providing a coherent coverage of the whole subject. In particular, minimal coverage is given to planetary magnetospheres, comets and to solar plasma physics, all of which now arguably come under the remit of space plasma physics. Furthermore, a more thorough discussion of the application of every process described would have been welcome through the inclusion of more experimental data. Two surprising omissions include the lack of coverage given to magnetic reconnection and to Birkeland (magnetic field-aligned) currents, surely processes

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

The diamond ring effect — a solar eclipse.

fundamental for driving plasma transport within the Earth's magnetosphere and in coupling solar wind momentum down to the ionosphere. More significantly, however, the textbook seems outdated, for it fails to convey either the true scope or the excitement of present-day research. Indeed one might argue that the same textbook could have been written almost a decade ago.

Physics of Space Plasmas is an attractive and worthwhile publication providing a timely, though far from exhaustive survey of a subject which must surely grow in importance during the 1990s. I recommend the book as the best textbook currently available for an undergraduate course and to all research groups interested in space plasma physics. But there remains a pressing need for an up-to-date scholarly text covering the whole of Solar System plasma physics, in which theory and observation are carefully balanced. □

Mark Saunders is in the Blackett Laboratory, Imperial College of Science, Technology and Medicine, London SW7 2BZ, UK.

Modelling the mind

L. H. Shaffer

Foundations of Cognitive Science. Edited by Michael Posner. MIT: 1990. Pp.888. \$50, £44.95.

COGNITIVE science has come into prominence in recent years as an innovative discipline for studying the human mind. Its central idea is that cognition is a species of computing and hence that the brain is a kind of computer. The requirements of a task and the mental processes in its performance are to be stated in computational terms. The computational description of a task, such as playing chess, carrying out medical diagnosis, or understanding speech, allows us to consider the class of possible minds that can perform it; whereas data from human performance allow us to converge on the biological mind within this class. Whether a model of mind, real or possible, is adequate for a given task can be tested by simulating it on a digital computer.

In different chapters of *Foundations of Cognitive Science* edited by Michael Posner, Zenon Pylyshyn discusses the computational concept of mind, whereas Allen Newell *et al.* and David Rumelhart discuss radically different architectures for the mental computer.

The aim of making the models of mental function explicit enough to be testable formally or by using computer simulation introduces a level of stringency lacking in the earlier theories of mind. The challenge of this stringency has already had remarkable results, rendering obsolete many of the traditional concepts inherited from philosophy, psychology and biology. Not the least benefit of the new science is that it exposes very clearly how far short the models still fall in accounting for human performance, sometimes providing clues of what is needed to improve them. This point tends to escape the critics of cognitive science, who seem to prefer the woolly comfort of bland generalization.

Posner is to be congratulated on having brought together so many distinguished authors, each contributing a tutorial chapter on their particular branch of the science. The best of these are exemplary and few of them fall below a high standard. The coverage of the science, even in a book this size, is by no means comprehensive, and that was not intended: there is enough to give the reader the flavour and some of the substance of the research, and each chapter contains a wealth of reference.

The chapters are organized in three sections: foundations, domains and assessment. But the two chapters in the last section, on philosophy of mind and cultural taxonomies, could well fit in the first, whereas the chapters on grammar and semantics in the first

would fit better in the second. The point is that foundations deals with general issues of how to conceive and study the mind, whereas domains deals with substantive areas of this study, and language is one such area. A reader not already familiar with cognitive science would be better off reading the latter section first to get a grasp of the research problems, before grappling with the more abstract ways of considering them.

Cognitive science is a meeting place of ideas from computer scientists, psychologists, linguists, philosophers, engineers, neuroscientists, and others interested in the design of intelligent machines, particularly of the machines instantiated by the human brain. In seeking to understand how a person performs a given task, we can start by trying to discover the class of mechanisms that can perform it, or by trying to infer the actual mechanism from performance data. The first tactic lends itself to the style of the computer scientist and engineer, the second tactic to that of the traditional experimental psychologist. The best of recent cognitive science combines these styles, constructing a continuous interaction between theory and data. But though it is not discussed, it becomes evident from reading the book that until recently the theoretical disciplines have made much of the running, producing a cornucopia of models that have left the psychologist flat footed trying to find experimental data to test them. The traditional methods of experimental psychology were simply too narrowly conceived to adapt to the newer ideas. Thus, some of the experimentally based chapters do not quite rise to the occasion, being too myopically concerned with secondary issues.

There is a penalty for this rather one-sided input to the science. Philosophers, computer scientists and linguists have tended to take a cartesian rationalist view of mind, emphasizing its ability to interpret its sensory input and use logical processes to reason and make decisions about an external world. This view follows fairly naturally from the computer metaphor. It is in marked contrast to the earlier cybernetic view of the organism as a control system regulating its world through feedback. The emphasis then was on the organism learning about its world by interacting with and manipulating it. This view also gave prominence to the fact that people in the real world are often confronted with ill-defined tasks and problems, and are often required to perform in a timescale that does not allow a rational reflection of options. The range of ill-defined tasks includes such activities as social interaction, artistic creation and scientific research. As soon as we consider these it becomes obvious that we need a theory of improvisation and innovation. This is likely to take a form that extends our current ideas on heuristics, shifting our focus from the search for rules and algorithms.

The signs of a return to the cybernetic view are already apparent in the recent robotics research, where the emphasis is on sen-

sory—motor action. A central assumption of the research is that brains form models and representations of their world as bases for action. This is a more dynamic conception than one which regards mental models as bases for interpreting the world. *Foundations of Cognitive Science* only occasionally picks up this newer trend. Its two chapters on action, postponed to near the end, provide excellent surveys of motor-control theory, but do not tackle the larger issues of mental representation. Hence they fail to counter-balance the cartesian bias of the chapters on, for instance, model-theoretic semantics and mental models. Thus, in a sense the book reflects a phase in the history of cognitive science that may be already passing. But we should be grateful for having it described so expertly. □

L. H. Shaffer is in the Department of Psychology, University of Exeter, Exeter EX4 4QG, UK.

Sacred cows

C. Bailyn

Theory of Neutron Star Magnetospheres. By F. Curtis Michel. *Chicago University Press: 1991. Pp.517. Hbk \$79.95, £63.95; pbk \$34.95, £27.95.*

It is commonplace in some circles to decry the impersonality of scientific writing. The abstract, dry and totally objective tone of most scientific work, it is claimed, contributes mightily to the double misapprehension that science is both infallible and boring. F. Curtis Michel's monograph *Theory of Neutron Star Magnetospheres*, a work which challenges not only the conventions of scientific writing, but some of the accepted truths of contemporary astrophysics, is most assuredly neither.

Magnetized neutron stars are thought to be the source of some of the more bizarre patterns of radiation observed in the sky. Among the phenomena commonly attributed to neutron stars are radio pulsars, X-ray bursters, and (by most astronomers) gamma-ray bursters. The surprising observational data on all of these objects has kept theorists off balance for over two decades. Recent discoveries reported in these pages include eclipsing pulsars, globular cluster pulsars, and possible ultra-soft X-ray sources; with the continued improvement in observational capacities at all wavelengths this flood of unexpected objects is likely to continue unabated. Needless to say, theorists have not been shy of offering explanations for all of these phenomena, and in many cases a consensus seems to have been reached. But the foundations on which these theoretical edifices are constructed are often surprisingly weak.

In *Theory of Neutron Star Magnetospheres*, Michel has taken it upon himself

to demonstrate these flaws. Virtually every assumption, hidden or otherwise, invoked in the study of radio pulsars is dissected, and in many cases rejected. I found myself alternately delighted at Michel's intellectual courage, and infuriated at his refusal to see reason, depending upon which sacred cow he was engaged in goring. Most researchers in these fields will experience both of these reactions in studying this book, probably in different places from me and from each other. Michel is a clever debater, and has a deep understanding of the rather complex physics underlying pulsar magnetospheres, and I found that the effort of trying to refute some of his seemingly implausible statements deepened my understanding of the theories, and in some cases changed my initially sceptical attitude. Arguing with this book is not a trivial exercise, but it is one which anybody who presumes to write research papers on pulsars should be required to undertake.

But the highly individualistic and argumentative tone of the book has some drawbacks. Chief among these is that it makes the book unsuitable for beginning graduate students, or for researchers in related fields who wish to acquire an overview of current work on neutron stars. This is unfortunate, because the field is advancing rapidly and there is currently a need for a book-length review. But the choice of topics is too idiosyncratic to fulfil this need: a work that dwells for dozens of pages on the speculation that Jupiter should be considered a radio pulsar, but devotes only a single paragraph to the considerable observational and theoretical work devoted to quasiperiodic oscillations in X-ray binaries, which are far more relevant to neutron-star magnetospheres, is hardly a place to begin one's study of the field. To take another example, the many pages devoted to skewering the hypothesis of magnetic-field decay in pulsars would seem pointless to anyone not already steeped in the standard arguments in its favour. The organizational structure is also weak, as Michel leaps between basic physics, phenomenology and observational data, often referring the reader to a crucial figure or argument hundreds of pages forwards in the book. And there are enough of the usual typographical errors and minor misstatements to annoy pedants and trap the unwary.

Despite these quibbles, one must be grateful to Michel for allowing us such unusually close access to his opinions and prejudices. If more scientists were as forthcoming, many fewer mistakes would be repeated and perpetuated. So I hope we will see more such works in the future, in astrophysics and in science generally, despite the treacherous terrain they would present for those not previously acquainted with their subject matter. □

C. Bailyn is in the Department of Astronomy, Yale University, PO Box 6666, New Haven, Connecticut 06511, USA

The early development of terrestrial ecosystems

William A. Shear

New work on the fossil record of early terrestrial ecosystems is challenging the validity of the accepted picture of their development. In particular, palaeoecologists are learning of the dangers of drawing analogies between Palaeozoic organisms and their modern counterparts, and are beginning to rely more on direct inferences from the fossils themselves.

THE first terrestrial ecosystems are now thought to have been based on plants of a moss-like grade of organization, mats of algae, crusts of lichens and cyanobacteria, or even on plants unlike any alive today. These mats, crusts and clumps were probably colonized by small, air-breathing arthropods long before the arrival of vascular plants. In other words, some of the animal components of terrestrial ecosystems may have been on land waiting for the major primary producers of today to arrive. By the time vascular plants appeared, their productivity was available to animals only after the plants (or parts of them) had died. Carnivores and detritivores, not herbivores, dominated terrestrial animal communities for tens of millions of years before animals that could eat the living tissues of higher plants appeared. During the late Palaeozoic, dominance at the level of primary production in terrestrial communities was evidently shared by a wider range of higher taxa than is now the case; the tree form was assumed by members of every major vascular plant clade.

The origins of terrestrial life

Examining the colonization of the land by single species^{1,2} illuminates the individual physiological adaptations developed in the process, but the pioneering of new habitats is always a function of a community of species. For this reason, the community will be the focus of this review.

Beerbower³ has discussed in detail the conditions that may have been present on the abiotic continents. He emphasized the difference in the rate of chemical weathering and the release of mineral nutrients in the absence of macrophytic plants. A practical absence of organic detritus would have lowered the partial pressure of CO₂ in ground water, reducing the production of carbonic acids. Fungi would have largely lacked a substrate, so even if present, they would have contributed to leaching only in a surface film⁴. Without leaf cover or root mats to slow the movement of water through the sediments, hydrological regimes would have been 'flashy', with heavy flooding during storms, and rapid drying afterwards⁵. The water in streams and lakes would have been low in mineral nutrients because of reduced chemical weathering, and would probably have carried a heavier load of suspended sediments. Lack of the binding effects of macrophyte cover would also have resulted in rapid physical erosion and instability of the surface. Temperatures at the surface and in the soil would have fluctuated violently with changes in insolation.

On a worldwide basis, the absence of evapotranspiration and a more uniform planetary albedo would have inhibited the transfer of atmospheric moisture, so that the interiors of continents would probably have had unstable rainfall regimes. But it seems that short-wave ultraviolet radiation, often cited as an environmental problem for early land life, would have been strongly reduced if the partial pressure of atmospheric oxygen were as little as 1% of the present value⁶. Long-wave ultraviolet, far less damaging to life, may still have presented problems in the early Palaeozoic.

The most favourable environments for the invasion of land would have been coastal lowlands³, not only because of their proximity to the sources of colonists, but because most of the factors mentioned above would be ameliorated by low relief and proximity to large bodies of water. Such habitats were common from the Cambrian onwards.

The earliest terrestrial communities were probably microbial crusts and mats⁷, which may have appeared on damp substrates

BOX 1 Divisions of geological time

Palaeozoic	Permian	Late	248 Myr	
		Early	258	
	Carboniferous	Pennsylvanian	Stephanian	286
			Westphalian	296
		Mississippian	Namurian	316
			Visean	333
		Devonian	Tournaisian	352
			Famennian	360
			Frasnian	367
			Givetian	374
	Eifelian		380	
	Emsian		387	
	Siegenian		394	
	Gedinnian		401	
	Silurian	Pridoli	408	
		Ludlow	414	
		Wenlock	421	
		Llandovery	428	
	Ordovician	Ashgill	438	
		Caradoc	448	
		Llandeilo	458	
		Llanvirn	468	
		Arenig	478	
		Tremadoc	488	
			505	

Guide to some of the divisions of geological time mentioned in the text. The subdivisions of the Ordovician and Silurian Periods are considered Epochs, while those of the Devonian are Ages. The Epochs of the Devonian are prosaically named Early (Gedinnian, Siegenian, and Emsian), Middle (Eifelian and Givetian) and Late (Frasnian and Famennian). The Namurian, Westphalian and Stephanian in the Carboniferous are each divided into A, B and C, going from oldest to youngest. The Subperiods Mississippian and Pennsylvanian are used primarily in North America. The dates are subject to periodic revision and are in millions of years ago. The Brigantian (also mentioned in the text) is at the top of the Visean (latest Visean).

even in the Precambrian, leaving as traces palaeosols enriched in organic matter⁸. Intertidal cyanobacteria would have been preadapted for the formation of these crusts, as they would already have been able to resist alternate wetting and drying, and fluctuating salinities. One well-studied extant microbial crust consists mainly of two species of cyanobacteria, one normally fresh-water and the other marine; this example occurs in a desert environment, a strong indication of the potential adaptability of such communities⁸. Only regions unsuited for colonization by vascular plants are available for microbial crusts today, but such communities may have been varied and widespread from Precambrian times⁹ until the appearance of macrophytes in the Ordovician or Silurian (see Box). Few well-studied fossil examples are available. Kobluck *et al.*¹⁰ studied a karstic surface in the earliest Gedinian (Devonian) in Ontario and suggested that algal, lichen, or moss cover was present. The oldest known ascomycete fungi were found in Ludlovian (Upper Silurian) rocks in Sweden¹¹. Ascomycetes are components of all but a small percentage of lichen symbioses; lichens are important pioneers of abiotic environments. The Swedish specimens also included what were interpreted as microarthropod faecal pellets, and 'animal-type' cuticles, indicating that early microbial/lichen communities, like their extant analogues, probably harboured populations of tiny herbivores, detritivores, and their predators³.

Development of an organic soil (humus) under such mats would have been extremely slow, especially if large burrowing animals, which mix surface organic material and underlying mineral particles, were absent⁷. Burrows have been described in an Ordovician palaeosol¹² which may be attributable to large annelids or arthropods; their activities would have aided in mixing and their faeces would have enriched the soil in available nitrates and phosphates¹³. In addition to their support of animal communities, early mats of microbes and/or lichens would also have accelerated chemical weathering of mineral substrates. Oxalic and other acids produced by lichens and fungi are of great importance in dissolving minerals from rock¹⁴. The activity of these communities occurring in already favourable habitats would have—in the absence of disturbance—set up a closed positive feedback loop in which increased biological activity would have made the site even more suitable for life. Eventually, the accumulation of organic soil components and a greater range of animal life may have created conditions favourable for invasion by, or *in situ* development of, more complex photosynthetic forms. These forms may have been early eukaryotic green algae, charophytes, or even their close relatives, the embryophytic plants.

Embryophytes (liverworts, hornworts, mosses and vascular plants) may have appeared as early as the middle Ordovician^{15–18}. The recent debate on the timing of the advent of such plants on land has been instructive. The old paradigm suggested that plants developed most of their terrestrial adaptations (conducting strands, cuticle, stomata, resistant spores, gametophyte and sporophyte life-cycle phases) synchronously and probably in an aquatic phase. The sudden appearance in the fossil record of vascular plants also engendered an emphasis on macrofossil evidence and a persistent confusion between the category 'land plants', an ecological grouping, and vascular plants, a morphologically defined taxon^{19,20}. The failure to find a conducting strand in early specimens of *Cooksonia*²¹, long regarded as the earliest vascular plant, opened the door to the more realistic model in which the main adaptations of land plants arose in at least two phases, if not in a mosaic fashion^{22,23}.

Gray¹⁶ has maintained that the initial preadaptation leading to the colonization of land was the development of the resistant spore, probably in a *Coelochaete*-like green alga. It now seems irrefutable that embryophytes sprang from such an ancestor, and that the vascular plants are paraphyletically nested within the group conventionally called the bryophytes, having the mosses as their closest relatives²⁴.

The earliest meiospores attributed by Gray to embryophytes

or embryophyte progenitors are found in Llanvirn–Llandeilo (Middle Ordovician) sediments in Arabia¹⁸. They consist of obligate tetrahedral tetrads of smooth-walled spores, often enclosed in a common outer wall probably contributed by the mother cell (Fig. 1). Such spores are very close in form to those produced by living primitive liverworts²⁵. The earliest macrofossils of liverworts are lowermost middle Devonian or late Silurian²⁶. The ecology of certain extant bryophytes suggests that Ordovician or early Silurian hepaticoid plants may have been able to withstand severe environmental stress by means of an ephemeral life style—colonizing a temporarily favourable habitat and maturing and producing propagules (resistant spores) before conditions deteriorated. Yet other bryophytes are extremely tolerant of desiccation, surviving for long periods of time in a dry state, to 'come back to life' when moisture is once again available²⁷. The production of tetrads of spores is an advantageous strategy for plants with unisexual gametophytes (which occur in living hepatics with obligate dispersed tetrads) and in which the sex of the gametophyte is controlled by a chromosomal mechanism. Each tetrad has the potential to produce two male and two female gametophytes, increasing chances for fertilization and providing some limit, however small, to inbreeding²⁸.

In the later early Silurian, single dispersed spores (Fig. 2) appear for the first time^{17,18}. These spores bear the characteristic triradial mark found only in meiospores with a resistant wall, and which are today known from less than 3% of living moss genera²⁹, and from all vascular plants except those that flower under water and have secondarily lost the resistant wall¹⁷. Extant homosporous vascular plants (that is, ferns) have gametophytes which are at least potentially hermaphroditic and self-fertilizing. The sporophytes of such plants achieve rapid colonization and domination of a pristine habitat by vegetative reproduction; inbreeding to preserve adaptive complexes allowing survival in restrictive habitats would be advantageous. But as density increases, the environment may become more stabilized. As competition between individuals increases, so might the advantages of outbreeding, which would occur as a natural correlate of a higher density of gametophytes. The combination of rapid vegetative reproduction and the potential for greater heterozygosity would have allowed early vascular plants to dominate most environments rapidly, perhaps even by the early Silurian, driving the hepatics and other bryophyte-grade plants into their present limited ecological niches^{15–17}. There was an increase in spore size during this interval, so that some of the single trilete spores found in the early Silurian are larger than earlier entire tetrads¹⁵. Small differences in spore size can make a great difference in dispersal potential, with large spores representing

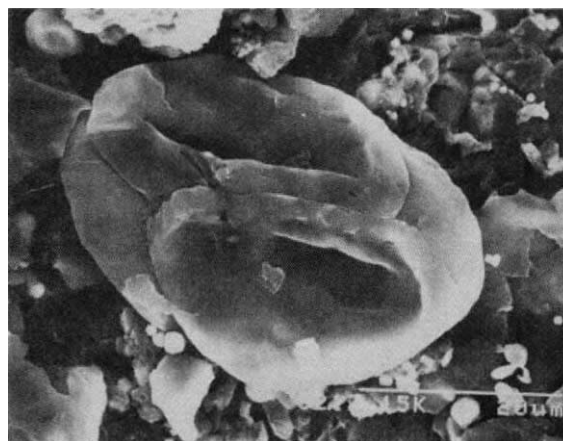


FIG. 1 Tetrad of meiospores from early Silurian rocks in Arabia. Scanning electron micrograph courtesy of Jane Gray.

an adaptation to limited dispersal³⁰, permitting a build-up of local populations in favourable habitats. With the origin of vascular tissue, organ differentiation in sporophytes was accelerated, particularly in the early and middle Devonian²². Around this time, vascular plants might also have formed mycorrhizal relationships that would have greatly increased their water and mineral absorbing capabilities³¹. Sporophyte-dominant lines took over the richest habitats; gametophyte-dominant plants remained in low-resource, stressful habitats.

Isolated scraps of peculiar, characteristic cuticle and both smooth and banded tubes that closely resemble vascular plant tracheids coincide at some sites with the spores Gray describes^{15,32,33}. The tracheid is the most characteristic feature of a vascular plant, being a tubular cell, non-living at functional maturity, whose walls are strengthened by thick spiral or circular bands containing lignin. The cuticles, which first appear in middle Ordovician spore macerations¹⁵, continue with little variation at least into the later middle Devonian. Several distinct types of cuticle have been recognized³³, and there is strong evidence³⁴ for association with enigmatic extinct plants known as nematophytes, whose bodies consisted of clusters of banded and smooth tubes, often of at least two distinct sizes. These plants seem to have been quite diverse, and included (in *Prototaxites*) the largest land organisms of their time, trunk-like structures up to a metre in diameter and several metres long^{35,36}. Quite unlike any extant plants but almost certainly terrestrial and photosynthetic, nematophytes were evidently important components of early land communities. Lignin or lignin-like compounds were isolated³⁷ from one of the earliest occurrences of nematophytes in the Llandovery (Silurian) of Virginia³⁸, where they are found with land plant spores. Such compounds have also been found in mosses and in the charophyte alga *Coelocochaete*, thought to be the closest relative of embryophytes³⁹. No *in situ* spores have as yet been associated with nematophytes.

Do at least some of Gray's spore tetrads, cuticles, and banded tubes derive from nematophytes, or are they evidence of early vascular plants as yet unknown from macrofossils? The earliest land plant macrofossil remains are the tiny (2–4 cm tall) fertile axes of *Cooksonia*, found in the Welsh Upper Wenlock (early in the late Silurian)⁴⁰. No tracheids have ever been demonstrated in these early finds, though short, sterile axes with tracheids co-occur with *Cooksonia* in later, Ludlow, sediments⁴¹. By late Silurian time, diverse assemblages of land plants, including probable vascular plants, bryophytes, hepatic-like thalli and nematophytes are found²⁶. This evident diversity is reflected in the dispersed spore record, with many new spore morphologies, including elaborations of the spore coat^{15,16}.

The adaptations of vascular plants probably arose only after their ancestors were subjected to the strong selection pressures of the terrestrial environment^{15,16,22}. The earliest of these pressures were nonbiological; only later did competition between plants become important, leading primarily to a 'stature race', the taller winners gaining more access to light. Probably many of the plant species present in the late Silurian were ecologically interchangeable, with little distinction between species in habitat preferences⁴². Diversity in local areas was low, despite increasing taxonomic diversity; vegetated areas consisted of dense stands of one or a few species of plants⁴³.

There is little evidence of animal life on land during this interval. Burrows from the Ordovician of Pennsylvania¹² remain a subject of debate. Could microbial mats or clumps of hepatics have provided a base for relatively large herbivores resident in burrows? But even the earliest microbial mats could have supported populations of animals occupying at least two higher trophic levels³. The coprolite evidence¹¹ from the late Silurian of Sweden is complemented by the scattered, rather poorly preserved fossils of millipedes and millipede-like animals from the Old Red Sandstone of Britain⁴⁴. Recently palaeozoologists have adopted some of the palaeobotanical techniques used to

isolate microfossils from shales—by dissolving the rock itself in hydrofluoric acid. Bits of plants and animals, consisting largely of highly reduced and polymerized organic matter, resist the acid and may be picked from the sample⁴⁴. A newly reported fauna, as yet not studied in detail, has been recovered in this way from Ludlow rocks (early Upper Silurian; about 420 Myr old) in Wales⁴⁵. This discovery is the oldest known for body fossils of terrestrial animals. Abundant animal remains had already been obtained by this means from a few Devonian localities (see below). Remarkably, the Ludlow fauna seems to share many evolutionarily advanced elements (trigonotarbid, centipedes) with the middle Devonian Gilboa fauna, 40–45 Myr younger.

Modernization in the Devonian

By the end of the Devonian, there existed terrestrial ecosystems that were at least architecturally modern⁴². Nearly all the main adaptations of vascular plants except flowers and fruit were complete, and all the main clades (lycopsids, sphenopsids, ferns, seed plants) had emerged^{22,26,42,47}. As abiotic selective factors continued to decline in importance in colonized habitats, competition between major clades of plants produced rapid architectural (and probably physiological) diversification. By the end of the Devonian, recognizable forests dominated by progymnosperm, lycopsid and sphenopsid trees, with a probable shrubby understorey and ground cover of ferns, herbaceous lycopsids and small seed plants, had appeared⁴². Competition would also lead to more sharply defined habitat preferences. Particular strategies seem, however, to have been characteristic of only one or a few species in each case, probably making late Devonian plant communities less stable than modern ones with greater 'guild depth'⁴². In addition, there is an emerging consensus regarding considerable biogeographic diversity in floras⁴⁷.

Body fossil evidence of animals in a community setting comes from early and middle Devonian localities in North America and Europe. The oldest of these—the Rhynie Chert of Scotland—is Pragian, 400–410 Myr old⁴⁸. As with all early to middle Devonian fossil faunas, it consists entirely of small arthropods. Much better understood and evidently far more diverse is the Givetian Gilboa fauna of New York, probably 30 Myr younger⁴⁵. The faunal assemblage at Gilboa is strongly predator-dominated, both in terms of individuals and species (Fig. 3). The few non-predator species present were probably detritivores, and there is no evidence for direct consumption of live plant material. The predator populations must have supported themselves on abundant detritivore species, as in today's litter faunas, but there is no obvious taphonomic reason why the remains of

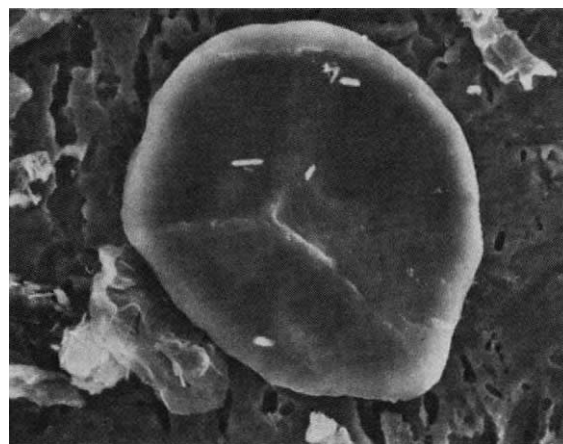


FIG. 2 Single dispersed spore with triradiate mark, isolated from Llandoveryan sediments in Pennsylvania, USA; about 430 Myr old. Scanning electron micrograph courtesy of Jane Gray.

the detritivores would not have been preserved. It may be that the hydrofluoric acid extraction selectively destroys some kinds of fossils, but this would be very hard to check. Although many long-extinct groups are represented, in functional terms the individual species are fully modern, and in a few cases it would be no surprise to find them among the extant litter fauna^{49–54}. Given classic assumptions about rates of evolution, the modern aspect of these animals and the occurrence of similar forms in the late Silurian Ludlow fauna⁴⁶ suggests that their ancestors had already spent a long period adapting to the land habitat, and leads one to predict that steadily older, similar faunas should be found, extending back at least into the early Silurian. If, however, adaptation and diversification were initially rapid, followed by essential stasis and selective winnowing, the earliest invasions may have taken place as late as the Middle Silurian.

It seems likely that transfer of primary productivity to higher trophic levels in early ecosystems based on algae, lichens and mosses or moss-like plants took place through microherbivores as well as detritivores. But with the dominance of vascular plants, the direct link between primary producers and consumers is very much weakened, if it does not disappear altogether⁴². A possible explanation for this decoupling of plant and animal worlds, which lasted even through the early Carboniferous, may lie in the structural molecule lignin, abundant only in vascular plants. The synthesis of this compound from phenolic precursors results in a number of toxic by-products, which, as plants are unable to excrete, must be stored in cell walls and dead tissues^{55,56}. These poisonous secondary compounds, coupled with the low nutritive value of fresh vegetative parts of plants, may have deterred animal herbivory for a long time. Although initially fortuitous, the production of these toxins by plants would have responded to natural selection through refinement of their toxic effects, and the development of new biosynthetic pathways producing unrelated toxins. Only after plant litter had been partially autolysed, and decomposed by bacteria and fungi,

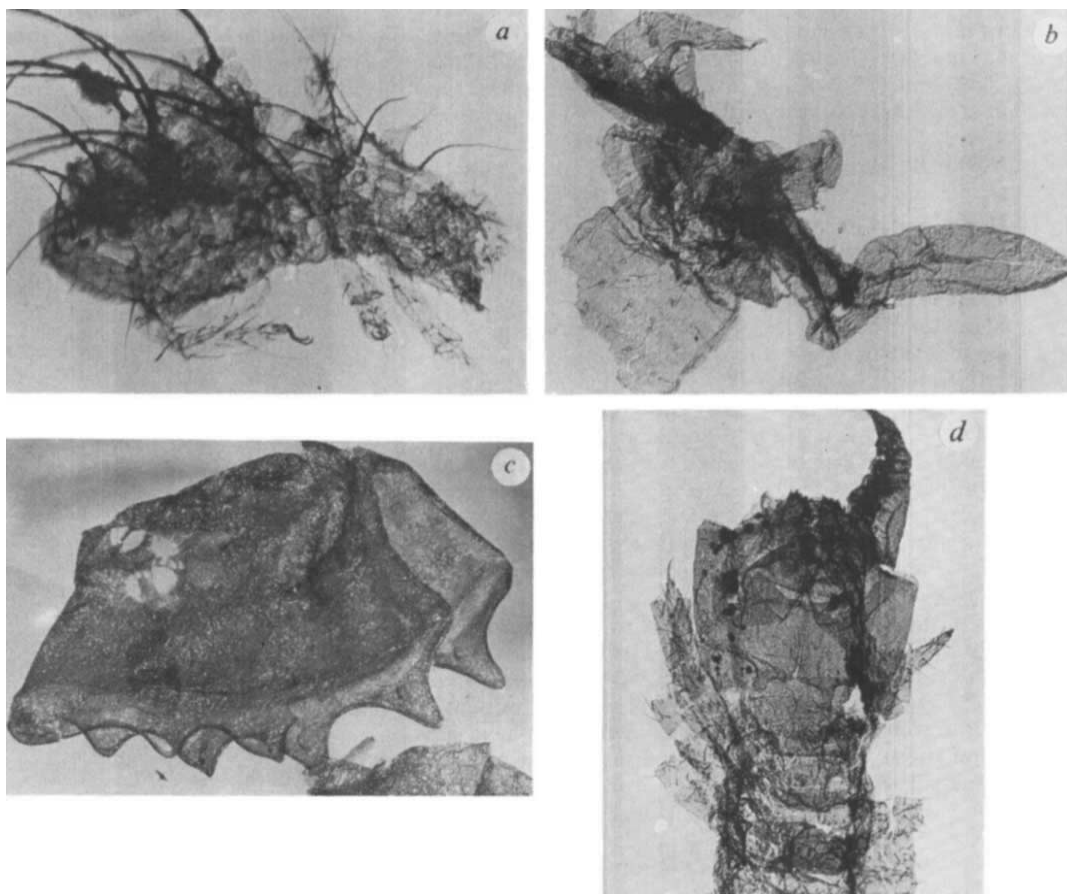
would it have been attacked by animals. The vegetative parts of plants are not nutrient-rich, being especially low in sodium and aromatic amino acids, and the complex carbohydrate polymers making up cell walls resist digestion by animals⁵⁵. Hence the early emphasis on detritivory—feeding on plant matter that had been partially autolysed and further broken down by microbes and fungi. The latter improve the nutritional quality of litter, and may have been the main foodstuff of many of the small arthropods¹¹.

Thus, the modernization of the architecture of plant communities was not matched by the development of plant–animal interactions, largely absent through much of the middle to early late Palaeozoic, nor by the evolution of animal food webs based on herbivory⁴². Among arthropods and vertebrates, two main strategies have made herbivory possible. The exploitation of microbial, protocistal or fungal mutualists, both internal and external, may have provided some detoxification and greatly enhanced the available calorific and nutritional value of plant material⁵⁶. Likewise, the consumption of more ephemeral plant parts such as spores, ovules and seeds, or young leaves, before they accumulated significant concentrations of toxins, was adopted early on by insects; the stored foods in reproductive structures makes them (at least in calorific terms) a much richer food source than vegetative parts⁵⁷. In addition, insects and some vertebrate herbivores possess enzyme complexes that act as general detoxification agents, and some species sequester plant toxins, even using them in their own defence^{55,56}.

Insects and tetrapods rise to dominance

Insects represent an extraordinarily successful line within the myriapod–hexapod assemblage. There is still no fossil evidence bearing on their origin. The sister group of the Insecta, Parainsecta (Collembola), first appears at Rhynie with two or more species so modern in aspect that they may be assignable to extant families⁵⁸; obviously the origin of this group and its

FIG. 3 Identifiable arthropod cuticles isolated from middle Devonian rocks near Gilboa, New York State. *a*, *Devonacarus sellnicki*, a primitive orbitatid mite. *b*, *Dracochela deprehendor*, the earliest known false scorpion. *c*, Carapace of *Gelasinotarbicus reticulatus*, a trigonotarbid arachnid, illuminated partially with incident light to show the slightly metallic lustre of the Gilboa fossils. *d*, *Devonobius delta*, the earliest known centipede and the type of a new order of centipedes.



dichotomy with Insecta lies in the more distant past. Parainsectans are adapted for life in confined spaces; their entognathous jaws are enclosed between the labrum (upper lip) and sidelobes of the head. Parainsectans initially used hiding as their main defensive adaptation, whereas true insects relied upon jumping and running to escape predators⁵⁷. Later, Insecta would split again along these lines and the entognath condition would arise once more independently. The first true insect, *Gaspea palaeoentognatha*, is known from a single specimen unaccompanied by any other terrestrial animal fossils. It occurred in palaeobotanical macerates of Lower Devonian rocks from Quebec⁵⁹. The lack of associated animal fossils, diagenetic changes, and ambiguities in the anatomical interpretation of *Gaspea* have led to doubts about its being a fossil⁴⁶. Remains of related animals, wingless archaeognathan machilids, occur at Gilboa⁴⁵. Following these occurrences, there is a long gap covering most of the Devonian and early Carboniferous until in the early Upper Carboniferous, winged insects seem suddenly to appear at a moderate, but rapidly increasing, level of diversity⁶⁰. But few localities are known; the Namurian B deposit at Hagen-Vorhalle, Germany, now shows all palaeopterous orders as well as diverse hemipteroids, suggesting that the 'explosion' of insect diversity in the late Carboniferous may only be apparent⁵⁷. Any increase in insect diversity in the late Carboniferous and early Permian must have been linked to the increasing diversity of vascular plants during the same period, though whether the numbers of available niches expanded gradually or remained essentially stable after a rapid, early diversification remains to be determined for insects and Palaeozoic plant life^{61,62}. Shear and Kukalová-Peck⁵⁷ favour the latter alternative; the palaeodictyopteroid orders remained stable through the Carboniferous and disappeared abruptly with the old Carboniferous flora of seed ferns, lycopod trees and cordaites.

Fossil evidence for Carboniferous insects feeding by chewing on living vegetative parts of plants is weak. Scott and Taylor⁶³ looked for reports of 'bite marks' on *Neuropteris* leaves, and examined collections in the Field Museum (Chicago). They concluded that such marks were "quite common", but when they examined a collection of 100 randomly chosen leaves from Pit 11 at Mazon Creek (late Carboniferous of Illinois), only 4 could be identified as bitten or chewed. The problem of demonstrating that the leaf, stem, or wood was alive or part of a living plant when attacked further complicates such interpretations. How does one differentiate a bitten dead leaf from a bitten living leaf on the basis of fossil evidence? In analysing stems, signs of healing or of reaction tissue are of aid. The same arguments apply to wood-boring (reported from the Carboniferous⁶⁴) though, as a dead tissue, wood cannot be expected to heal. Leaf-mining, known only from a few fossil examples, may be more reliable^{62,65}.

Among the earliest winged insects, members of the palaeodictyopteroid orders had specialized mouthparts used for tearing apart loosely organized, primitive cones (Fig. 4), or piercing ovules and sucking up the contents; their guts are sometimes crammed with spores⁶⁶ (Fig. 5). These short-lived plant parts would have been nutritious and low in toxins. A more complete analysis of insect mouthpart evolution, recently completed by Labandeira⁶², may shed more light on the early development of insect herbivory, but for the time being it seems unlikely that feeding on the leafy parts of plants, at least by early Carboniferous insects, was widespread. But there is convincing indirect evidence that many Palaeozoic hemipteroids fed by sucking: triangular stylet-like or bristle-like mouthparts and a pronounced, swollen pumping apparatus on the front of the head^{57,62}. Perhaps the original foods of these groups included plant cell contents and semi-liquid decaying matter. If phloem was a target, they would have required a mechanism to eliminate the excess water.

Pollination, another insect-plant interaction, may also have originated in the Carboniferous^{57,62,67}, though the evidence is

circumstantial; medullosan seed fern pollen, for example, is often so large that wind pollination is not likely⁶⁸. Megasecopteran and diaphanopteran insects also bore structures which, among other functions, may have served to collect and disperse pollen⁵⁷.

Predatory insects of the Carboniferous were typified by the often gigantic Protodonata, relatives of today's dragonflies. They were probably specialists in plucking other large insects from the trunks of lycopod trees in the coal swamps. Unable to fold their wings, many groups of Carboniferous insects would not have been able to walk through tangled vegetation and indeed would have been restricted to the open structure of the coal swamp 'pole-forest'. These predators and their prey were probably participants in a size-based arms race, as prey evolved to avoid predation by becoming larger, matched by increases in size on the part of the predator. By the mid-late Carboniferous, the largest insects that ever lived exhibited wingspans of more than 60 cm⁵⁷.

The early evolution of communities of tetrapods is analogous to that of arthropods. The earliest evidence from body fossils appears in the late Devonian of Greenland, Australia, and the western Soviet Union⁶⁸. The skeletons of *Ichthyostega* and *Acanthostega* from Greenland are already quite modern⁶⁹, but show some characters that might exclude them from the ancestry of extant tetrapods^{70,71}. *Acanthostega* in particular appears to have been entirely aquatic⁷². Ecologically, many of the diverse lines of amphibians that had appeared by the early Carboniferous, some of them very large animals, bridged aquatic and terrestrial environments. More terrestrially adapted forms (anthracosaurs and temnospondyls) also appeared very early. All the late Devonian and early Carboniferous tetrapods were predatory, feeding on fish, one another, and, most important for terrestrial ecosystems, insects. No potentially herbivorous tetrapods appear in the fossil record until the Stephanian (latest late Carboniferous), when functional morphology of the jaws of edaphosaurs (pelycosaurid reptiles) and diadectid amphibians suggests possible adaptations for feeding on vegetation, though large, hard seeds or shelled invertebrates would have been other possibilities. By the early Permian, the caseid pelycosaurs had developed an obvious suite of dental and other adaptations for feeding on vegetation⁷³.



FIG. 4 Reconstruction of *Homaloteneura lehmani* feeding on a *Cordaites* cone. The wing colour pattern, frequently preserved in Palaeozoic insect fossils, may have been disruptive or served for intraspecific signalling. Drawing courtesy of Jarmila Kukalová-Peck.

Undisputed amniotes (reptilomorphs) first appear in the Lower Carboniferous (Brigantian) of East Kirkton, Scotland⁷³; there is insufficient evidence to allow them to be placed in trophic relationships, but later (Westphalian) forms were insectivores. Finally, at the beginning of the Upper Permian, the dicynodont therapsids show undisputed adaptations for herbivory⁷³.

Thus the early tetrapods, like the earliest arthropods, were predators. Some of them retained extensive adaptations to life in water, and probably fed there, on fish or aquatic invertebrates. The more terrestrial forms were insectivores, so that detritivorous arthropods and potentially herbivorous insects remained, through much of the late Devonian and Carboniferous periods, the mechanism through which primary productivity by plants reached higher trophic levels. Gigantism among myriapods and a few other ground-dwelling arthropod taxa may have been an escalating response to tetrapod predation⁷⁵.

Of the two groups, the insects doubtless had the greater effect on the evolution of plant life. A wider range of relationships was possible because of the diversity of insect feeding strategies^{62,63} and their probable early specialization on fructifications⁵⁷. Even in the late Devonian, closed strobili, hardened seed coats, and protected micropyles had evolved to deter insect predation⁷⁶. Most herbivorous vertebrates were and have remained relatively unspecialized as to diet, and consequently have had little evolutionary impact on plant diversification⁷⁷.

Late Palaeozoic diversity and disparity

An examination of the diversity of palaeocommunities has become a standard exercise^{77,79}, but often without the required caution as to the taxonomic level at which the diversity is assessed, and without taking into account the varying levels to which the taxonomy of extinct groups has developed^{78,82}. In any case, systematic diversity is not synonymous with ecological diversity; perhaps a more meaningful concept is the recognition of ecomorphotypes^{80,81}. As adaptive groups emerge, they may be compared and contrasted with traditional taxonomic measures of diversity. Gould⁸³ and Briggs⁸⁴ have suggested that the range of body plan types and ecological strategies could be referred to as the 'disparity' of the community.

Although such studies are still infrequent in the emerging

subdiscipline of Palaeozoic terrestrial ecology, what has been done, and what can be inferred from the taxonomic literature, implies strong differences in community organization in the late Palaeozoic as compared to extant ecosystems⁴². Carboniferous coal swamp communities have been quantitatively studied through surveys of remains preserved in coal balls. DiMichele, Phillips and colleagues⁸⁵⁻⁹¹, have found that whereas coal swamp communities were characterized by low species-level and ecomorphotype diversity, at high systematic levels an extraordinary range of plants were present. The tree form was assumed by species from every major group of vascular plants—lycophytes, seed ferns, sphenopsids, ferns and gymnosperms—in contrast to forest communities of today, where one major taxon (angiosperm seed plants) dominates but shows extraordinary ecomorphotypic diversity^{42,80}. Beginning in the Permian, this unusually high taxonomic diversity within ecomorphotypes began its decline⁴², and when angiosperms first appear in the Cretaceous, they rapidly come to dominate the vegetation at nearly every level^{76,92}.

Candidates for a similar analysis would be the insects and tetrapods of the Carboniferous. In insects, for example, 10 extant orders and stem groups of nearly all of the others were present by the early Permian, in addition to several now extinct⁶². An extraordinary variety of amphibian and reptilomorph types has also been documented among the tetrapods⁷³.

Other possible interpretations

We cannot, however, be complacent about our revised views of Palaeozoic communities. For example, the cases of two enigmatic plant fossils, *Baragwanathia* and *Longfengshania*, suggest that it may be necessary to re-evaluate the accepted view of land plant evolution, even if Gray's spore evidence from the Ordovician continues to gain wide acceptance. *Baragwanathia* is a large, advanced fossil lycopod, with well developed leaves, an advanced type of vascular tissue, and sporangia in the leaf axils. Related plants are found abundantly in the Lower Devonian, but *Baragwanathia* occurs anomalously in the Upper Silurian of Australia, well before any other so highly evolved plants appear in the Northern Hemisphere²². While there is continuing debate about the dating of the rocks enclosing the *Baragwanathia* flora, it is still possible that vascular plants might have evolved first in Gondwana and reached the present Northern Hemisphere continents by a series of south to north migrations. The controversy over *Baragwanathia* emphasizes the importance of finding more fossil plant sites in Silurian rocks in all parts of the world.

Zhongying Zhang suggested that an abundant but enigmatic 880-Myr-old fossil from the Precambrian, *Longfengshania*, has many features in common with extant mosses and liverworts, as well as generally accepted Silurian and Devonian bryophyte fossils like *Torticaulis* and *Sporogonites*⁹³. This fossil has previously been interpreted as algal, but Zhang's arguments raise new questions about the tetrads of trilete spores obtained by Volkova from the late Precambrian of Latvia. Gray and Boucot⁹⁴ have used Volkova's evidence to suggest an earlier, failed, Precambrian origin and radiation of land plants. New fossils or new interpretations of previously discovered fossils may significantly alter our views of both the timing and geography of terrestrialization.

Similarly, lycopod trees are perhaps the best known of all extinct plants. Yet they were entirely unlike the angiosperm trees of today, having more resemblance to giant weeds⁴². They warn us that generalizations drawn from modern life may give only sketchy or misleading clues about the life of the past, which must be approached on its own terms.

Future research directions

The most urgent need in Palaeozoic ecological studies today is for more hard data in the form of detailed descriptions of fossils. As with Devonian plants, the repeated hashing over of the same

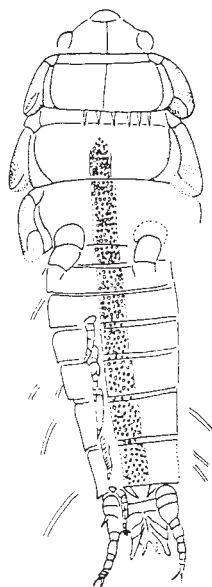


FIG. 5 Many Upper Carboniferous insects fed from plant fructifications or other ephemeral tissues, perhaps to avoid toxins, and because spores were relatively high in nutrients. Nymph of a fossil diaphanopteroid insect from the Upper Carboniferous of Illinois; the gut is packed with spores. Drawing courtesy of Jarmila Kukalová-Peck.

body of data has reached a point of diminishing returns. To carry this example further, far more information is required on the Devonian floras of the Southern Hemisphere and ancestral Gondwana⁹⁵, where the presence of the enigmatic *Baragwanathia* hints at the development of advanced floras long before their appearance in the north. Studies are now under way that should aid in defining the elements of early terrestrial arthropod communities, as well as pushing back the earliest dates of occurrence into the middle and early Silurian. As yet, the discovery of few of these communities has been the result of deliberate searching in appropriate deposits; such examination should be rewarding.

As data accumulates, qualitative analyses elucidating adaptive complexes of species, community organization, and the structure of food webs can begin. DiMichele, Phillips and colleagues have pointed to the next step: where possible, quantitative examinations of community composition and the correlation of this information with different geological terrains and time horizons. Important generalizations, such as the impact of abiotic stress on ancient plant communities, have already emerged from their work⁹⁶. □

William A. Shear is in the Department of Biology, Hampden-Sydney College, Hampden-Sydney, Virginia 23943, USA.

1. Little, C. *The Colonisation of Land: Origins and Adaptations of Terrestrial Animals* (Cambridge University Press, 1983).
2. Little, C. *The Terrestrial Invasion: an Ecophysiological Approach to the Origins of Land Animals* (Cambridge University Press, 1990).
3. Beerbower, R. in *Geological Factors and the Evolution of Plants* (ed. Tiffney, B. H.) 47–91 (Yale University Press, New Haven, 1985).
4. Kubiena, W. L. *Micromorphological Features of Soil Geography* (Rutgers University Press, New Brunswick, 1970).
5. Schumm, S. A. *Geol. Soc. Am. Bull.* **79**, 1573–1588 (1968).
6. Ratner, M. I. & Walker, J. C. G. *J. Atmos. Sci.* **29**, 803–808 (1972).
7. Wright, V. P. *Phil. Trans. R. Soc. B* **309**, 143–145 (1985).
8. Campbell, S. E. *Origins of Life* **9**, 335–348 (1979).
9. Golubic, S. & Campbell, S. E. *Precamb. Res.* **8**, 201–217 (1979).
10. Kobluk, D. R., Pemberton, S. G., Karolyi, M. & Risk, M. J. *Bull. Can. Petrol. Geol.* **25**, 1157–1186 (1977).
11. Sherwood-Pike, M. & Gray, J. *Lethaia* **18**, 1–20 (1985).
12. Retallack, G. & Feakes, C. *Science* **235**, 61–63 (1987).
13. Jones, C. G. & Shachak, M. *Nature* **346**, 839–841 (1990).
14. Wilson, M. J. & Jones, D. *Geol. Soc. Spec. Publ.* **11**, 5–12 (1983).
15. Gray, J. *Spec. Pap. Geol.* **32**, 281–295 (1985).
16. Gray, J. *Phil. Trans. R. Soc. B* **309**, 167–192 (1985).
17. Gray, J. & Boucot, A. *Lethaia* **10**, 145–174 (1977).
18. Gray, J., Massa, D. & Boucot, A. *Geology* **10**, 197–201 (1982).
19. Axelrod, D. I. *Evolution* **13**, 264–275 (1959).
20. Banks, H. P. in *Evolution and Environment* (ed. Drake, E. T.) 73–107 (Yale University Press, New Haven, 1968).
21. Edwards, D. & Fanning, U. *Phil. Trans. R. Soc. B* **309**, 147–165 (1985).
22. Knoll, A. H., Grant, S. W. F. & Tsao, J. W. in *Land Plants* (ed. Gastaldo, R. A.) 45–63 (University of Tennessee, Knoxville, 1986).
23. Edwards, D., Fanning, U. & Richardson, J. B. *Nature* **323**, 438–440 (1986).
24. Mishler, B. D. & Churchill, S. P. *Cladistics* **1**, 305–328 (1985).
25. Clarke, G. C. S. *Syst. Assoc. Spec. Vol.* **14**, 231–250 (1979).
26. Taylor, T. N. *Taxon* **37**, 805–833 (1988).
27. Scott, G. A. in *Bryophyte Ecology* (ed. Smith, A. J. E.) 105–122 (Chapman & Hall, London, 1982).
28. Wyatt, R. J. *Hattori Bot. Lab.* **52**, 179–198 (1982).
29. Erdtman, G. *An Introduction to Palynology, III* (Almqvist & Wiksell, Stockholm, 1965).
30. Mogensen, G. S. *Bryologist* **84**, 187–207 (1981).
31. Stubblefield, S. & Taylor, T. N. *New Phytol.* **108**, 3–25 (1980).
32. Edwards, D., & Rose, V. *Bot. J. Linn. Soc.* **88**, 35–54 (1984).
33. Edwards, D. *Bot. J. Linn. Soc.* **84**, 223–256 (1982).
34. Strother, P. K. *J. Paleontol.* **62**, 967–982 (1988).
35. Niklas, K. J. *Rev. Palaeobot. Palynol.* **22**, 1–17 (1976).
36. Schmid, R. *Science* **191**, 287–288 (1976).
37. Niklas, K. J. & Pratt, L. M. *Science* **209**, 396–398 (1980).
38. Pratt, L. M., Phillips, T. L. & Dennison, J. M. *Rev. Palaeobot. Palynol.* **25**, 121–149 (1978).
39. Delwiche, C. F., Graham, L. E. & Thomson, N. *Science* **245**, 399–401 (1989).
40. Edwards, D. & Feehan, J. *Nature* **287**, 41–42 (1980).
41. Edwards, D. & Davies, E. C. W. *Nature* **263**, 494–495 (1976).
42. DiMichele, W. A. et al. in *Evolutionary Paleocology of Terrestrial Plants and Animals* (University of Chicago Press, in the press).
43. Gensel, P. G. & Andrews, H. N. *Am. Scient.* **75**, 478–489 (1987).
44. Almond, J. *Phil. Trans. R. Soc. B* **309**, 227–237 (1985).
45. Shear, W. A. et al. *Science* **224**, 492–494 (1984).
46. Jeram, A., Selden, P. A. & Edwards, D. *Science* **250**, 658–661 (1990).
47. Scheckler, S. E. in *Land Plants* (ed. Gastaldo, R. A.) 81–96 (University of Tennessee, Knoxville, 1986).
48. Hirst, S. *Ann. Mag. Nat. Hist.* **9**, 455–474 (1923).
49. Kethley, J. B., Norton, R. A., Bonamo, P. M. & Shear, W. A. *Micropaleontology* **35**, 367–373 (1989).
50. Norton, R. A., Bonamo, P. M., Grierson, J. G. & Shear, W. A. *J. Paleontol.* **62**, 259–269 (1988).
51. Shear, W. A. & Bonamo, P. M. *Am. Mus. Novit.* **2927**, 1–30 (1988).
52. Shear, W. A., Palmer, J. M., Coddington, J. A. & Bonamo, P. M. *Science* **246**, 479–481 (1989).
53. Shear, W. A., Schawaller, W. & Bonamo, P. M. *Nature* **341**, 527–529 (1989).
54. Shear, W. A., Selden, P. A., Rolfe, W. D. I., Bonamo, P. M. & Grierson, J. D. *Am. Mus. Novit.* **2901**, 1–74 (1984).
55. Weis, A. E. & Berenbaum, M. R. in *Plant–Animal Interactions* (ed. Abrahamson, W. G.) 123–162 (McGraw-Hill, New York, 1989).
56. Lindroth, R. L. in *Plant–Animal Interactions* (ed. Abrahamson, W. G.) 163–206 (McGraw-Hill, New York, 1989).
57. Shear, W. A. & Kukalová-Peck, J. *Can. J. Zool.* **68**, 1807–1834 (1990).
58. Greenslade, P. & Whalley, P. E. S. *2nd Internat. Symp. Apterygota* **1**, 319–323 (1986).
59. Labandeira, C. C., Beall, B. S. & Hueber, F. M. *Science* **242**, 913–916 (1988).
60. Wootton, R. J. A. *Rev. Ent.* **26**, 319–344 (1981).
61. Strong, D. R., Lawton, J. H. & Southwood, R. *Insects on Plants* (Harvard University Press, Cambridge, 1984).
62. Labandeira, C. C. & Beall, B. S. in *Arthropod Paleobiology* (ed. Mikulic, D.) 214–256 (1990).
63. Scott, A. C. & Taylor, T. N. *Bot. Rev.* **49**, 259–307 (1983).
64. Cichan, M. A. & Taylor, T. N. *Palaeogeogr. Palaeoecol.* **39**, 123–127 (1982).
65. Rozefelds, A. C. *Proc. R. Soc. Queensl.* **99**, 77–81 (1988).
66. Kukalová-Peck, J. *Can. J. Zool.* **63**, 933–955 (1985).
67. Crepet, W. L. *BioScience* **29**, 102–108 (1979).
68. Milner, A. R., Smithson, A. C., Milner, M. I., Coates, M. I. & Rolfe, W. D. I. *Mod. Geol.* **10**, 1–28 (1986).
69. Clack, J. A. *Palaeontology* **31**, 699–724 (1988).
70. Panchen, A. L. & Smithson, T. R. in *The Phylogeny and Classification of the Tetrapods, vol. 1: Amphibians, Reptiles, Birds* (ed. Benton, M. J.) 1–32 (Clarendon, Oxford, 1988).
71. Coates, M. I. & Clack, J. A. *Nature* **347**, 66–69 (1990).
72. Bendix-Almgreen, S. E., Clack, J. A. & Olsen, H. *Terra Nova* **2**, 131–137 (1990).
73. Carroll, R. L. *Vertebrate Paleontology and Evolution* (Freeman, New York, 1988).
74. Smithson, T. R. *Nature* **342**, 676–678 (1989).
75. Vermeij, G. *Evolution and Escalation* (Princeton University Press, New Haven, 1987).
76. Taylor, T. N. *Paleobotany* (McGraw-Hill, New York, 1981).
77. Wing, S. L. & Tiffney, B. H. *Rev. Palaeobot. Palynol.* **50**, 179–210 (1987).
78. Niklas, K. J., Tiffney, B. H. & Knoll, A. H. *Evol. Biol.* **12**, 1–89 (1980).
79. Niklas, K. J. in *Patterns and Processes in the History of Life* (eds Raup, D. M. & Jablonski, D.) 383–405 (Springer, Berlin, 1986).
80. Wing, S. L. et al. in *Evolutionary Paleocology of Terrestrial Plants and Animals* (University of Chicago Press, in the press).
81. Damuth, J. in *Evolutionary Paleocology of Terrestrial Plants and Animals* (University of Chicago Press, in the press).
82. Smith, A. B. & Patterson, C. *Evol. Biol.* **23**, 127–216 (1988).
83. Gould, S. J. *Wonderful Life* (Norton, New York, 1989).
84. Briggs, D. E. G. in *Arthropod Paleobiology* (ed. Mikulic, D.) 24–43 (University of Tennessee Press, Knoxville, 1990).
85. DiMichele, W. A. & Phillips, T. L. *Rev. Palaeobot. Palynol.* **44**, 1–26 (1985).
86. DiMichele, W. A. & DeMaris, P. J. *Palaios* **2**, 146–157 (1987).
87. DiMichele, W. A. & Phillips, T. L. *Rev. Palaeobot. Palynol.* **52**, 115–132 (1988).
88. DiMichele, W. A., Phillips, T. L. & Peppers, R. A. in *Geological Factors and the Evolution of Plants* (ed. Tiffney, B. H.) 223–256 (Yale University Press, New Haven, 1985).
89. Phillips, T. L. *Rev. Palaeobot. Palynol.* **27**, 239–289 (1979).
90. Phillips, T. L., Peppers, R. A. & DiMichele, W. A. *Int. J. Coal Geol.* **5**, 43–109 (1985).
91. Phillips, T. L., Peppers, R. A., Acvin, M. J. & Laughan, P. J. *Science* **184**, 1367–1369 (1974).
92. Fredrickson, N. D. *Geosci. Man* **4**, 17–28 (1972).
93. Zhang, Z. *28th Internat. Congr. Geol. Abst.* **3**, 441 (1989).
94. Gray, J. & Boucot, A. *Geology* **6**, 489–492 (1978).
95. White, M. E. *Am. Scient.* **78**, 252–262 (1990).
96. DiMichele, W. A., Phillips, T. L. & Olmstead, R. G. *Rev. Palaeobot. Palynol.* **50**, 151–178 (1986).

ACKNOWLEDGEMENTS. My collaboration with P. Bonamo provided the impetus for this review. I also thank J. Gray, P. Selden, J. Kukalová-Peck, P. Gensel, J. A. Clack and W. D. I. Rolfe for discussion. W. DiMichele, R. Gastaldo, R. Beerbower, J. Boy, N. Hotton, R. Hook, T. Phillips, S. Scheckler, and H.-D. Sues participated in the Evolution of Terrestrial Ecosystems Conference at Airlie, Virginia, in 1987, and with the author are contributors to chapters in the forthcoming book, *Evolutionary Paleocology of Terrestrial Plants and Animals*. W. A. DiMichele permitted citation of this book, where many of the syntheses and ideas used in this review were first developed at length. This paper was written while my research was supported by the NSF and the Power Authority of the State of New York.

Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules

Kirsten Falk, Olaf Rötzschke, Stefan Stevanović*, Günther Jung* & Hans-Georg Rammensee†

Max-Planck-Institut für Biologie, Abteilung Immunogenetik, Corrensstrasse 42, W-7400 Tübingen, Germany

* Institut für Organische Chemie, Universität Tübingen, Auf der Morgenstelle 18, W-7400 Tübingen, Germany

The crystal structures of major histocompatibility complex (MHC) molecules contain a groove occupied by heterogeneous material thought to represent peptides central to immune recognition, although until now relatively little characterization of the peptides has been possible. Exact information about the contents of MHC grooves is now provided. Moreover, each MHC class I allele has its individual rules to which peptides presented in the groove adhere.

T LYMPHOCYTES recognize their antigens in context of MHC-encoded molecules, a phenomenon called MHC restriction¹⁻⁵. Crystallography of human MHC class I molecules, HLA-A2 and Aw68, revealed a groove made up by the $\alpha 1$ and $\alpha 2$ domains of heavy chains^{3,6}. This groove is believed to be the binding site for antigenic peptides, as both crystals contain peptide-sized structures not compatible with the MHC sequences located at that groove⁶. T cells can recognize synthetic peptides loaded on MHC class I molecules, and MHC-associated peptides representing T-cell epitopes have been extracted from normal or virus-infected cells^{2,4,5,7,8}. Similarly, the antigens recognized by MHC class II-restricted T cells can be mimicked by artificial peptides⁹, and MHC-associated antigenic peptides have been eluted from MHC class II molecules¹⁰. By virtue of their position in the middle of trimolecular complexes made up of T-cell receptor, peptide and MHC molecule¹¹, T-cell epitopes are literally central to the specific immune system and the so far unknown rules defining them should be central to understanding it¹²⁻¹⁵.

Comparison of the first naturally processed H-2K^d-restricted T-cell epitope⁴ with several peptides that contain other K^d-restricted epitopes revealed some common features not found in non-K^d-restricted epitopes⁴⁷. Assuming that all naturally processed K^d-restricted epitopes are nonapeptides, all those peptides containing K^d-restricted epitopes could be aligned with a Tyr residue at the second position and an amino-acid residue with a side-chain methyl group (Val, Ile, Thr, Ala or Leu) at the last position, suggesting a K^d-specific peptide motif. (Tyr is important in K^d-restricted peptides, and some allele-specificity of T-cell epitopes has been noticed before¹²⁻¹⁶.) Sequencing the self-peptide blends eluted from purified K^d molecules should reveal this K^d-specific motif. If so, sequencing of peptide mixtures from other MHC class I molecules should reveal the respective motifs if they exist.

Elution of peptides from K^d molecules

K^d molecules were immunoprecipitated from detergent extracts of P815 tumour cells (H-2^d). Bound peptides were dissociated from bead-associated K^d molecules by acid treatment⁷ and separated by reversed-phase HPLC (Fig. 1a, b). Precipitation from influenza-infected cells (not shown) gave a K^d-restricted

influenza epitope in fraction 24, indicating that the method can isolate MHC-bound peptides. Heterogeneous material elutes in low amounts between fractions 20 and 28 (Fig. 1a, b), covering the range of many MHC class I-restricted T-cell epitopes with this HPLC-gradient (refs 4, 7, 8, and O.R. *et al.*, manuscript submitted).

K^d-restricted peptide motif

Fractions 20-28 were pooled both from the K^d-derived batch and from a mock precipitate, and both batches sequenced automatically using the Edman degradation method (Table 1). This method involves the sequential derivatization and removal of amino acids from the N terminus, each of which is identified chromatographically. Because it is unusual to sequence complex mixtures of peptides, we have presented the raw data from the sequencer. Table 1a and b shows the results from two sequencing attempts for K^d-eluted peptides. Table 1c shows the sequencing result of a mock elution with D^b-specific antibodies on P815 lysates. The K^d-eluted peptides have a distinct amino-acid residue pattern for each position from 1 to 9, whereas the mock-eluted material shows a uniform pattern of residues throughout, with a decrease of the absolute amount of each residue with every cycle. Thus, for the K^d-eluted peptides, only residues showing more than 50% increase in the absolute amount compared with the previous or the pre-previous cycle were arbitrarily considered significant and are underlined. The first position is difficult to judge; there is no previous cycle, and all free amino acids present in the HPLC pool are detected in this position. For the second position, the only residue whose frequency is clearly increased by comparison with the previous cycle is Tyr, in both attempts (Table 1a, b), for example, 60.9 to 875.6 pmol. The only other residue that shows an increase, however marginal, is Phe, which has a side chain similar to that of Tyr. This confirms our premises resulting from comparing the natural K^d-restricted influenza epitope TYQRTALV (single-letter amino-acid code) with other K^d-restricted peptides, with regard to the Tyr at position 2. By contrast to the second position, there is no single amino-acid residue standing out in the following positions up to 8, although up to 14 different residues are detected in the individual positions. At position 9, the residues detected are Ile and Leu. This again agrees with our premises. There is no signal increase at position 10, indicating that most K^d-bound self-peptides are no longer than 9 residues. The natural K^d-restricted influenza peptide is also a nonapeptide⁴. The consensus sequence pattern indicated by these data is shown in Table 1d. Most evident are Tyr at position 2 and Ile or Leu at 9, whereas at all other positions a larger, but distinct set of residues was found. A comparison of this motif with peptide sequences containing K^d-restricted epitopes indicates that, if aligned by their Tyr residues, most fit quite well with the consensus K^d-restricted nonamer motif (Table 1d).

Sequence of a prominent self-peptide

The peak marked by an arrow in fraction 29 of Fig. 1b and the corresponding fraction of the mock precipitation were rechromatographed giving higher resolution (Fig. 1c). The sharp specific peak was determined to be SYFPEITHI by direct

† To whom correspondence should be addressed.

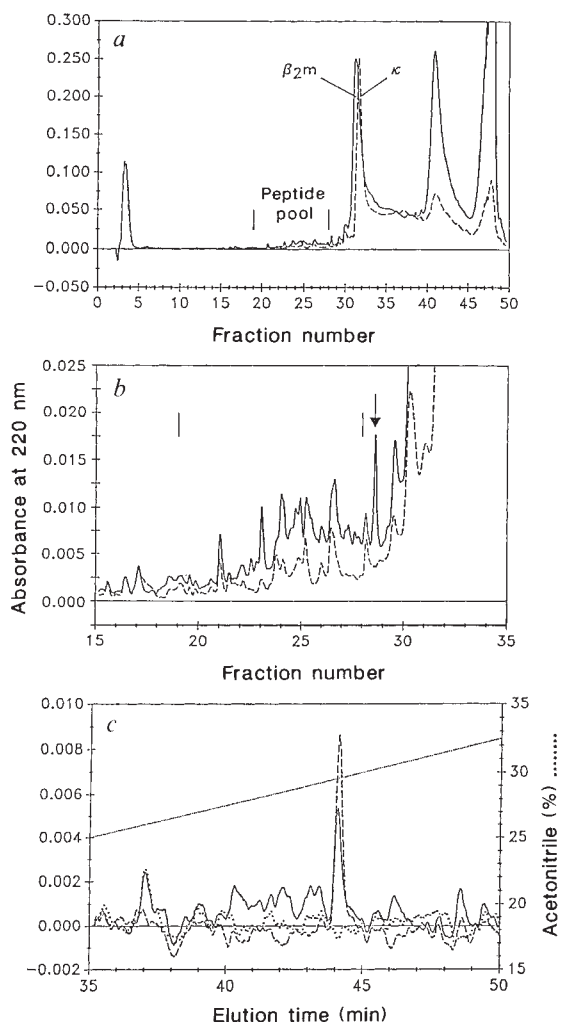


FIG. 1 HPLC separation of immunoprecipitated and TFA-treated K^d molecules. *a*, HPLC profile of TFA-treated material precipitated from P815 lysate with anti- K^d beads (—) or mock-precipitated with irrelevant anti- D^b beads (---). The material eluting beyond fraction 32 is probably MHC light and heavy chains and antibody components. The peak at fraction 48 is detergent. *b*, Enlarged view of the chromatogram in *a* (fraction 15–35). Note several distinct peaks in the K^d precipitate (—), presumably self peptides, which do not appear in the control precipitate (---). The most dominant peak is marked by an arrow (fraction 29) and was determined by sequencing to be SYFPEITHI. *c*, Rechromatography of the dominant self-peptide marked by an arrow in *b* (—), of the corresponding fraction in a control precipitate produced with glycine-coupled beads (·····), and of 500 ng synthetic SYFPEITHI (---).

METHODS. $10\text{--}20 \times 10^9$ P815 cells were pelleted and stirred for 30 min with 250 ml 0.5% Nonidet P-40 in phosphate buffered saline (PBS) containing 0.1 mM PMSF at 4 °C. The supernatant (after centrifugation at 250g, 5 min, followed by 150,000g, 30 min at 4 °C) was passed through a chromatography column (bed volume, 1 ml) filled with glycine-coupled beads, then through a similar column filled with anti- K^d beads, then for a mock-precipitate over anti- D^b beads. Beads removed from all three columns were swirled with 0.1% TFA for 15 min. Supernatants were dried by vacuum centrifugation and separated by reversed-phase HPLC using a Suprapac Pep S Column (C2/C18; 5 μ m particles, 4.0×250 mm; Pharmacia LKB) and Pharmacia LKB equipment⁴. Eluents: solution A, 0.1% TFA in H_2O (v/v); solution B, 0.1% TFA in acetonitrile. Gradient used in *a* and *b*: 0–5 min, 100% A; 5–40 min, linear increase to 60% B; 40–45 min, 60% B; 45–50 min, decrease to 0% B; flow rate 1 ml min⁻¹, fraction size 1 ml. Fraction size in *c*, 0.5 ml. Individual fractions were collected and dried by vacuum centrifugation. Antibody-coupled and glycine-coupled beads were prepared using cyanogen bromide-activated Sepharose 4B (Pharmacia LKB) according to manufacturer's protocol. K^d -specific antibodies (5 mg each) 20-8-4S (IgG2a, κ ; ref. 25) or D^b specific B22-249 (IgG2a, κ ; ref. 26) were coupled to 1 ml of beads. β 2m, β 2-microglobulin.

sequencing. Identity of this natural self-peptide was confirmed by its coelution with synthetic SYFPEITHI on HPLC (Fig. 1c). The sequence fits well with the motif from the pool of fractions 20–28 (Fig. 1a, b), thus confirming the K^d -restricted peptide motif (Table 1d). It will be of interest to see whether alloreactive T cells can recognize this prominent self-peptide, which we estimate, on the basis of comparison with the peak area of the coeluting synthetic peptide, to occupy about 5% of K^d molecules of P815 cells (10,000 SYFPEITHI molecules per cell).

Elution of peptides from K^b and D^b

Detergent lysates from EL4 tumour cells ($H\text{--}2^b$) were immunoprecipitated with K^b -specific and D^b -specific antibodies. Peptides dissociated from MHC molecules were separated by reversed-phase HPLC. Both K^b - and D^b -derived material eluted with profiles roughly similar to the K^d -derived material, with marked differences, however, in the heterogeneous material eluting between fractions 20 and 28 (not shown).

D^b -restricted peptide motif

Pooled fractions 20–28 from the D^b preparation were sequenced (Table 2a, b). Each of positions 2 to 4 contained several residues. By contrast, cycle 5 gave a strong signal for Asn, which was much weaker for cycle 4 (from 4.2 to 271.4 pmol in the first and from 6.7 to 154.7 pmol in the second experiment). Thus, the dominant residue at position 5 of D^b -eluted self-peptides is Asn. The weak signal for Asp is caused by hydrolysis of Asn to Asp under sequencing conditions. Positions 6–8 contained 5–14 different detectable residues. Position 9 contained a strong signal for Met, an intermediate one for Ile, and a weak one for Leu (all hydrophobic). (The importance of Met or Ile in a D^b -restricted epitope has been reported earlier¹⁷.) Position 10 had no signal, indicating D^b -presented self-peptides to be nonapeptides, as is the D^b -restricted influenza peptide⁴. The consensus motif indicated by these data is shown in Table 2c. Comparing this motif with the natural D^b -restricted peptide and with other peptides containing D^b -restricted epitopes shows that Asn at position 5 may be an invariant anchor residue (for definition, see Table 1) of the D^b -restricted peptide motif, not unlike Tyr at position 2 for the K^d -restricted motif. The other residues of the D^b -restricted epitopes vary considerably, with the exception of the Met, and also either Ile or Leu at position 9, which looks like a second anchor position. This anchor is not at position 9 in the (few) known T cell epitopes, except in the naturally processed one.

K^b -restricted peptide motif

Pooled fractions 20–28 from the K^b preparation were sequenced (Table 3a, b). No strong signal for any residue was at the second position. Position 3 contained a good signal for Tyr, and a weak one for Pro. Position 4 revealed weak signals for five residues. Strong signals for Phe (increase from 1.8 to 50.5 and from 1.5 to 18.3 pmol) and for Tyr make these two residues dominant at position 5. The next two positions contained five or three residue signals. Position 8 had a strong signal for Leu, an intermediate one for Met, and weak ones for Ile and Val. Position 9 showed no increase for any residue, consistent with the length of the known K^b -restricted natural peptide, which is an octamer⁵. Analysis of the consensus K^b restricted motif and comparison with epitopes indicates two anchor positions: Tyr or Phe (both with similar aromatic side chains) at position 5 and Leu, Met, Ile or Val (all with similar hydrophobic side chains) at position 8.

HLA-A2.1 restricted peptide motif

Detergent lysate of human JY cells (HLA-A2.1) was immunoprecipitated with A2-specific antibodies. Peptides dissociated from A2 molecules were separated by HPLC. Fractions 20–28 were pooled and sequenced as for the mouse material (Table 4). The second position contained a strong signal for

TABLE 2 Sequencing of the self-peptide mixture eluted from D^b molecules*

(a) Experiment 1		Amino-acid residues (in pmol)																	
	A	R	N	D	E	Q	G	H	I	L	K	M	F	P	S	T	Y	V	
Cycle	Ala	Arg	Asn	Asp	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Tyr	Val	
1	257.2	18.2	21.6	1.3	8.1	16.3	99.1	2.3	22.0	21.2	20.3	7.2	33.0	27.5	124.6	43.9	26.9	70.1	
2	202.1	7.2	5.4	<u>6.8</u>	7.4	<u>24.7</u>	116.2	0.9	5.4	9.9	6.5	<u>154.1</u>	4.3	8.2	52.7	15.0	5.5	16.0	
3	29.9	5.9	5.3	0.0	3.8	5.5	<u>185.1</u>	1.1	<u>106.3</u>	<u>65.8</u>	0.0	8.3	3.8	<u>88.1</u>	8.3	4.7	5.2	<u>73.2</u>	
4	18.3	8.1	4.2	<u>4.6</u>	<u>32.4</u>	<u>21.8</u>	49.3	0.8	32.7	21.5	<u>12.4</u>	3.6	2.3	28.8	9.9	<u>18.6</u>	5.0	<u>165.2</u>	
5	6.8	2.1	<u>271.4</u>	26.0	8.2	4.3	43.0	0.6	4.7	6.2	2.5	1.3	0.9	11.7	4.5	5.0	1.7	7.6	
6	<u>42.1</u>	5.9	29.6	7.1	8.4	<u>7.8</u>	32.6	<u>1.3</u>	<u>18.0</u>	<u>148.4</u>	<u>8.8</u>	1.9	<u>11.3</u>	<u>22.5</u>	<u>7.8</u>	<u>11.8</u>	<u>4.1</u>	<u>23.6</u>	
7	21.5	23.4	18.2	<u>24.5</u>	<u>30.4</u>	<u>13.7</u>	22.0	0.7	9.9	16.2	2.4	<u>2.1</u>	3.6	16.4	6.7	<u>54.3</u>	<u>5.1</u>	<u>35.0</u>	
8	14.6	10.1	11.3	9.8	23.2	10.3	18.2	0.3	3.0	10.1	<u>4.4</u>	1.3	5.0	9.5	<u>26.5</u>	24.9	<u>12.5</u>	20.7	
9	7.5	3.2	7.9	3.2	3.1	1.6	11.2	0.5	<u>8.5</u>	13.7	0.5	<u>7.7</u>	3.0	2.5	2.0	3.3	3.6	3.5	
10	2.6	1.1	2.5	2.4	1.9	1.2	12.5	0.3	4.2	8.5	0.4	2.7	1.8	2.1	1.6	1.7	1.9	1.3	
(b) Experiment 2																			
1	413.4	45.8	29.7	15.9	14.5	19.6	132.4	4.7	41.5	40.8	48.9	17.2	50.8	26.1	307.7	94.0	47.4	110.1	
2	227.4	14.4	7.6	9.3	11.1	25.2	133.8	2.1	8.2	14.5	13.3	<u>169.9</u>	5.6	4.9	71.0	21.6	11.3	22.6	
3	39.6	3.3	6.0	6.3	6.0	5.3	<u>172.2</u>	1.2	<u>89.5</u>	<u>56.0</u>	1.6	14.7	4.5	<u>75.4</u>	12.1	5.0	7.6	<u>79.2</u>	
4	29.3	16.6	6.7	<u>10.6</u>	<u>34.8</u>	<u>23.0</u>	57.3	0.8	36.3	21.7	<u>17.0</u>	8.1	4.2	33.5	12.5	<u>23.9</u>	7.4	<u>198.9</u>	
5	19.9	5.3	<u>154.7</u>	22.2	8.7	4.1	31.1	0.9	4.6	7.0	4.3	2.4	1.7	11.8	5.3	5.0	2.0	13.8	
6	<u>42.3</u>	8.4	30.8	15.7	<u>14.6</u>	<u>8.3</u>	28.7	2.3	<u>18.6</u>	<u>124.1</u>	8.2	<u>5.3</u>	<u>11.2</u>	<u>22.1</u>	<u>7.9</u>	<u>10.7</u>	<u>5.6</u>	<u>29.2</u>	
7	22.0	24.5	15.4	<u>33.5</u>	<u>29.2</u>	<u>10.5</u>	17.7	1.6	11.3	14.8	3.3	3.7	3.6	14.3	7.5	<u>47.3</u>	<u>6.9</u>	<u>35.5</u>	
8	15.8	10.9	10.2	20.9	25.6	8.0	12.6	3.2	3.3	13.6	4.3	2.8	5.1	8.7	20.8	19.3	<u>12.9</u>	23.6	
9	8.7	4.3	6.1	13.0	12.1	2.6	8.7	0.3	<u>19.8</u>	<u>26.2</u>	<u>1.2</u>	<u>30.8</u>	3.9	4.4	4.8	5.6	7.2	9.2	
10	5.4	3.1	3.9	12.2	8.1	2.0	8.2	0.0	10.1	13.9	0.7	11.6	3.2	3.4	3.0	3.0	7.3	5.9	

(c) The D^b-restricted peptide motif

	1	2	3	4	5	6	7	8	9†	
	Position									
Dominant anchor residues					N				M	
Strong		M	I	K		L			I	
			L	E		F				
			P	Q						
			V	V						
Weak	A	A	G	D		A	D	F	L	
	N	Q		T		Y	E	H		
	I	D				T	Q	K		
	F					V	V	S		
	P					M	T	Y		
	S					E	Y			
	T					Q				
	V					H				
						I				
						K				
						P				
						S				
Known epitopes, aligned‡	A	S	N	E	N	M	E	T	M	
	S	G	P	S	N	T	P	P	E	/§
	S	G	V	E	N	P	G	G	Y	C
	S	A	I	N	N	Y	.	.		

Protein source

Ref.

Influenza NP366-374		4, 2
Adenovirus E1A		38
Lymphocyte choriomeningitis virus GP 272-293		39
Simian virus 40 T 193-211		40

* Lysates of EL4 cells were prepared as indicated for the P815 lysate in Fig. 1, and then passed through chromatography columns containing glycine beads, anti-K^b (K9-178, IgG2a, κ ; ref. 27) or anti-D^b beads. Anti-D^b beads were treated with 0.1% TFA, and the supernatant was separated by HPLC as for the K^d precipitate in Fig. 1. Fractions 20–28 were sequenced and analysed as in Table 1.

† Anchor positions indicated in bold types.

‡ Alignment was on the Asn residue at position 5. The epitope known to be naturally processed is underlined. For the other peptides, only the parts likely to be the naturally processed epitopes are indicated, although most of the peptides were described as longer ones in the original references.

§ Residues at position 10 and 11 are in italics. These peptides might be exceptional D^B-restricted epitopes having their C-terminal anchor residue at position 10 or 11 instead of 9. Note that both epitopes contain four 'small' amino-acid residues, Glv and Pro, which may influence the linear distance between N and C termini.

|| With the alignment proposed, positions 7–9 are not covered by the peptide published.

Leu and an intermediate one for Met. Positions 3-5 had 6-8 residues each. Position 6 contained Val, Leu, Ile, and Thr. Each of the following two positions had three signals. Position 9 had a strong Val and a weak Leu signal. Position 10 showed no increase for any residue, indicating A2-restricted epitopes to be nonapeptides. Anchors appear to be Leu or Met at position 2 and Val or Leu at position 9. Some of the peptides reported to contain A2-restricted epitopes can be aligned to the motif, whereas others can be aligned only partially (Table 4c). The existence of variant A2 molecules typing as A2 by serology may cause the poor alignment to the motif of some of the peptides.

The epitope content of some of these peptides has also not been formally established.

Discussion

Sequencing of the self-peptide mixtures from the four MHC class I molecules H-2K^d, H-2K^b, H-2D^b and HLA-A2 indicated a distinct allele-specific peptide motif presented by each class I molecule. K^d-, D^b- and A2-presented peptides are nonamers, whereas K^b-presented peptides seem to be octamers; the corresponding peptide motifs contain two anchor positions occupied by a fixed residue or by one of a few residues with closely related

TABLE 3 Sequencing of the self-peptide mixture eluted from K^b molecules*

(a) Experiment 1			Amino-acid residues (in pmol)																
Cycle	A	R	N	D	E	Q	G	H	I	L	K	M	F	P	S	T	Y	V	
	Ala	Arg	Asn	Asp	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Tyr	Val	
1	978.7	26.3	49.2	55.8	39.0	23.1	514.9	20.9	167.5	167.2	189.8	50.3	116.7	118.2	120.8	365.2	136.0	352.5	
2	345.5	3.9	37.3	41.8	23.5	20.3	475.2	8.9	44.5	43.1	72.6	12.6	25.4	51.0	253.1	80.5	50.1	93.5	
3	129.0	1.4	14.7	37.0	17.7	9.8	358.8	5.9	8.2	19.0	26.9	4.1	6.0	32.5	56.2	20.0	<u>75.6</u>	25.9	
4	52.1	<u>3.5</u>	10.8	45.3	<u>38.0</u>	9.2	246.7	5.0	4.9	7.0	17.7	2.4	1.8	14.6	23.0	13.4	12.0	16.4	
5	18.9	1.3	5.5	34.7	12.0	3.6	128.2	2.8	1.9	4.7	3.8	1.6	<u>50.5</u>	6.7	8.9	4.6	<u>33.2</u>	4.9	
6	16.2	0.8	5.6	32.7	13.0	3.7	77.9	2.4	<u>3.1</u>	3.5	3.9	0.9	4.5	7.3	9.2	<u>18.3</u>	7.3	6.2	
7	9.9	0.9	<u>14.9</u>	30.4	9.5	<u>6.6</u>	51.3	0.6	0.0	3.4	<u>9.2</u>	0.5	1.9	4.7	6.1	10.7	3.5	3.4	
8	6.0	1.4	5.1	22.7	6.0	3.3	29.2	0.8	<u>1.4</u>	<u>13.5</u>	1.8	<u>2.1</u>	1.0	3.8	4.1	3.1	2.5	3.6	
9	4.6	1.5	2.6	19.9	4.5	2.3	21.1	0.9	0.9	6.9	1.0	1.5	1.0	3.0	3.7	2.2	1.9	2.1	
10	3.9	0.5	1.9	17.5	3.7	2.1	17.5	1.0	0.5	4.0	0.8	0.9	1.2	2.8	3.5	1.8	2.0	1.5	
(b) Experiment 2																			
1	42.4	1.1	5.2	3.0	7.8	17.1	44.6	0.3	11.3	12.6	12.1	3.8	6.2	7.6	44.2	18.1	6.8	26.2	
2	24.0	0.2	<u>9.4</u>	2.8	5.1	8.0	42.5	0.5	4.7	6.3	4.0	1.3	3.7	3.5	14.9	10.3	3.1	6.9	
3	10.4	0.3	2.1	2.6	3.9	4.0	25.1	0.7	2.8	7.9	2.1	0.9	3.6	<u>9.8</u>	3.0	3.3	<u>16.7</u>	10.0	
4	9.6	<u>1.3</u>	2.7	<u>5.7</u>	<u>7.5</u>	4.1	24.5	0.2	1.5	5.0	<u>6.3</u>	0.7	1.5	5.9	3.0	<u>5.9</u>	2.7	4.5	
5	5.8	0.8	1.8	2.8	3.3	2.5	14.2	0.5	0.2	3.9	1.7	0.4	<u>18.3</u>	3.5	1.3	2.0	<u>20.8</u>	2.2	
6	8.6	0.2	2.3	2.7	<u>6.3</u>	2.7	9.2	0.0	<u>1.0</u>	2.4	1.5	0.4	2.3	3.2	<u>2.7</u>	<u>5.2</u>	3.6	2.4	
7	5.0	0.1	<u>8.2</u>	3.3	3.9	<u>4.2</u>	10.4	0.6	0.4	2.3	<u>7.2</u>	0.1	1.2	2.1	1.9	2.8	1.9	1.2	
8	4.0	0.1	3.1	2.0	2.6	<u>1.7</u>	6.9	0.2	0.2	<u>13.8</u>	1.6	<u>1.0</u>	0.8	1.1	0.7	1.3	1.1	<u>2.2</u>	
9	4.5	0.1	1.1	2.1	3.6	1.9	5.9	<u>1.4</u>	0.0	<u>7.7</u>	0.9	1.0	0.9	1.3	0.3	1.3	0.8	1.7	
10	3.9	1.7	0.3	4.5	3.0	1.4	5.4	0.2	0.0	3.9	0.6	0.6	0.6	1.1	0.6	1.1	0.8	1.1	
(c) The K ^b -restricted peptide motif																			
	Position																		
	1	2	3	4	S	6	7	8 [†]											
Dominant anchor residues					F			L											
					Y														
Strong			Y					M											
Weak		R	N	P	R		T	N	I										
		I			D		I	Q	V										
		L			E		E	K											
		S			K		S												
		A			T														
Protein source																			
Ref.																			
Known epitopes, aligned‡	R	G	Y	V	Y	Q	G	L	Vesicular stomatitis virus NP 52-59										5
	S	I	I	N	F	E	K	L	Ovalbumin 258–276§										41
	A	P	G	N	Y	P	A	L	Sendai Virus NP 321-332										42

* The anti-K^b beads loaded with EL4 lysates described in Table 2 were treated with 0.1% TFA as were the anti-K^d beads in Fig. 1. Extracted material was HPLC-separated as before, and fractions 20-28 were pooled, sequenced, and analysed as in Table 1.

† Anchor positions are in bold types.

‡ Alignment was on the Leu residue at position 8. The only known naturally processed epitope is underlined. For the others, only the postulated natural epitopes are indicated; additional residues contained in the original references have been omitted.

§ The published ovalbumin peptide did not contain the Ser residue at position 1.

side chains. These anchor positions are not at the same place in the different motifs; they are at position 5 and 9 (D^b), or 2 and 9 (K^d, A2), or 5 and 8 (K^b). The C-terminal anchor residues of all motifs are hydrophobic. For H-2L^d (data not shown), one anchor residue was Pro at position 2. (The entire motif could not be sequenced, as not enough material eluted from L^d molecules.) The residues not at anchor positions can be fairly variable; some, however, seem to be preferentially occupied; for example, Pro is prominent at position 4 of the K^d motif, Tyr at position 3 of the K^b motif, and hydrophobic residues predominate at positions 3 of the D^b motif and 6 of the A2 motif. Because of technical limitations of our unconventional approach, less abundant amino-acid residues may have escaped detection; Cys and Trp residues were not detected at all. Comparison of the motifs with known epitopes indicates that some residues have been missed by sequencing the self-peptide mixtures. A minor population of peptides not adhering to the motif-specific lengths also cannot be excluded. The K^d- and D^b-restricted epitopes whose positions 9 do not fit with the hydrophobic anchor of the respective motifs would have residues at position 10 or 11 that would fit (Tables 1d, 2c).

Our observations match well the structure of the peptide-binding cleft of MHC class I molecules^{3,6}. With HLA-A2, the

cleft has pockets of a size that would accommodate specific amino-acid side chains, for example Leu^{3,6}. Leu dominantly occupies position 2 of the A2 motif. Another pocket should exist in A2 molecules to accommodate the side chains of Val and Leu at position 9 of A2-restricted epitopes. Co-crystallizing material not from the A2 sequence and bound to the cleft showed extensions (possibly Leu and Val side chains) fitting the A2 pockets^{3,6}. With the Aw68 crystal, which differs from A2 at 13 amino-acid residues, pockets of different location and of different shape from that of A2 were found⁶. Therefore, different MHC class I alleles differ in the location and shape of pockets in the cleft likely to be able specifically to accommodate certain amino-acid side chains.

Thus, the allele-specific pockets in MHC crystals and the side chains of the allele-specific anchor residues we find in the sequenced self-peptides are likely to reflect complementary structures. From this, we deduce that the structure of K^d (mouse MHC molecules have not been crystallized yet) possesses a pocket for the aromatic side chains of Tyr or Phe at one end of the cleft and another one at the opposite end, able to accommodate Ile, Leu, or Val (all hydrophobic with side-chain methyl groups). Similarly, K^b should have a pocket close to the cleft's centre for the aromatic side chains of Tyr or Phe, and another

TABLE 4 Sequencing of the self-peptide mixture eluted from A2.1 molecules*

(a) Experiment 1			Amino-acid residues (pmol)																
Cycle	A	R	N	D	E	Q	G	H	I	L	K	M	F	P	S	T	Y	V	
	Ala	Arg	Asn	Asp	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Tyr	Val	
1	172.6	0.0	31.9	25.7	44.8	125.9	112.4	2.8	144.4	123.8	60.0	30.7	63.3	117.9	75.9	49.0	50.3	104.9	
2	42.5	0.0	16.2	14.1	25.6	53.1	44.7	1.6	69.6	<u>511.0</u>	15.5	<u>71.0</u>	10.5	38.7	16.2	16.1	12.2	86.5	
3	<u>99.8</u>	0.0	9.5	18.3	12.3	20.4	31.8	11.1	51.5	110.8	5.8	55.7	<u>19.4</u>	30.4	12.0	8.7	<u>20.9</u>	46.0	
4	36.0	0.6	12.7	<u>26.4</u>	<u>59.5</u>	21.7	<u>56.2</u>	1.3	10.4	22.7	<u>24.6</u>	5.2	5.2	<u>52.4</u>	10.9	<u>14.0</u>	5.2	28.8	
5	35.1	0.1	13.4	18.6	28.1	19.8	55.6	<u>2.8</u>	21.4	23.9	<u>47.2</u>	4.1	6.2	39.1	7.5	10.5	<u>11.6</u>	29.0	
6	30.3	0.9	16.8	14.1	21.4	17.3	28.5	1.4	<u>68.1</u>	<u>43.4</u>	14.7	4.4	5.8	40.8	9.2	<u>20.3</u>	5.0	<u>106.2</u>	
7	42.1	0.0	11.7	9.5	27.2	21.8	19.0	<u>3.2</u>	36.3	27.3	7.9	5.7	8.0	54.1	5.4	13.6	<u>14.0</u>	62.8	
8	37.9	0.3	13.4	8.1	<u>37.3</u>	24.3	21.1	1.8	11.6	15.1	<u>33.8</u>	3.4	5.1	22.3	<u>8.8</u>	17.9	10.2	22.4	
9	23.3	0.0	5.1	6.0	15.7	10.5	14.8	0.7	11.5	<u>27.5</u>	8.7	3.1	2.7	11.9	5.6	6.7	5.1	<u>60.2</u>	
10	12.0	0.7	2.6	4.4	6.5	5.2	10.2	0.4	4.5	12.1	4.5	1.0	1.8	7.1	2.7	3.2	2.3	20.4	
(b) Experiment 2																			
1	110.8	10.8	4.0	3.1	10.0	14.5	55.7	0.2	60.3	44.4	10.8	8.2	37.5	20.3	27.4	14.6	19.8	48.0	
2	13.4	1.6	2.0	1.9	6.8	11.0	9.0	0.0	37.9	<u>302.7</u>	0.0	<u>26.2</u>	5.0	6.3	4.4	4.5	3.3	26.5	
3	<u>62.4</u>	<u>3.5</u>	<u>5.0</u>	<u>9.1</u>	4.9	10.0	12.6	0.1	35.7	71.5	0.0	24.5	<u>13.8</u>	<u>13.4</u>	<u>8.9</u>	4.8	<u>17.9</u>	19.6	
4	16.9	2.2	4.5	8.9	<u>25.3</u>	7.9	<u>24.5</u>	0.1	6.2	10.3	<u>2.8</u>	1.3	2.0	<u>22.1</u>	4.9	5.0	1.8	9.3	
5	22.3	1.6	6.8	8.6	14.3	9.9	<u>31.8</u>	0.0	<u>16.6</u>	15.1	<u>8.2</u>	1.9	4.0	16.3	4.5	4.6	<u>5.7</u>	<u>18.3</u>	
6	10.6	1.3	6.6	3.6	6.4	6.2	10.1	0.1	<u>38.7</u>	<u>27.1</u>	0.0	1.4	2.7	12.6	3.2	6.1	1.3	<u>39.2</u>	
7	<u>19.2</u>	1.0	4.7	2.5	7.2	9.0	5.6	0.2	<u>22.3</u>	16.1	0.0	1.9	3.9	17.4	1.9	3.5	<u>3.6</u>	27.2	
8	13.4	1.2	3.1	1.3	7.9	6.3	6.9	0.3	4.7	6.7	<u>3.0</u>	0.6	2.0	5.1	2.2	4.9	1.6	5.3	
9	5.7	0.5	0.9	0.8	2.9	2.0	2.7	0.2	3.8	<u>11.5</u>	0.4	0.3	0.6	2.0	1.0	1.1	0.4	<u>10.8</u>	
10	2.9	0.6	0.5	0.5	1.0	0.9	1.8	0.3	1.6	5.1	0.4	0.3	0.3	0.8	0.4	0.3	0.2	3.6	
(c) The HLA-A2-restricted peptide motif																			
	1	2	3	4	5	6	7	8	9†										
Dominant anchor residue		L							V										
Strong		M		E		V		K											
				K															
Weak	I		A	G	I	I	A	E	L										
	L		Y	P	K	L	Y	S											
	F		F	D	Y	T	H												
	K		P	T	N														
	M		M		G														
	Y		S		F														
	V		R		V														
					H														
Reported epitopes, aligned‡	I	L	K	E	P	V	H	G	V										
	G	I	L	G	F	V	F	T	L										
	I	L	G	F	V	F	T	L	T	V¶									
	F	L	Q	S	R	P	E	P	T										
	A	M	Q	M	L	K	E												
	P	I	A	P	G	Q	M	R	E										
	Q	M	K	D	C	T	E	R	Q										
										Protein source								Ref.	
										HIV reverse transcriptase 461–485								43	
										Influenza matrix protein 57–68								44	
										Influenza matrix protein 57–68								44	
										HIV Gag protein 446–460§								46	
										HIV Gag protein 193–203§								46	
										HIV Gag protein 219–233§								46	
										HIV Gag protein 418–443§								46	

* Detergent lysate of JY cells (HLA-A2.1)⁴⁵ prepared as indicated for PB15 cells in Fig. 1 was passed through a column containing glycyl beads, then through a column containing anti-A2 (BB7.2, IgG2b; ref. 28) beads. Loaded anti-A2 beads were treated with 0.1% TFA, and the material in the supernatant was separated by HPLC as for the K^d preparation in Fig. 1. Fractions 20 to 28 were pooled and sequenced as in Table 1.

† Anchor positions are in bold types.

‡ Alignment was on the anchor at position 2 of the motif. Only the nonapeptides fitting to the motif are indicated, although all of the peptides were described as longer ones in the original references. For the influenza matrix peptide, an alternative alignment is shown, which employs an Ile residue at position 2. Ile and Leu have very similar side chains. With this alignment, position 9 of the peptide fits with the C-terminal anchor of the motif. In addition, position 6 of the peptide then fits well with position 6 of the motif, which may represent an auxiliary anchor of this motif preferentially for hydrophobic residues.

§ All four HIV Gag peptides can only partially be aligned with the motif. These peptides induced low or intermediate lysis of target cells at micromolar concentrations⁴⁶, whereas natural epitopes induce lysis at picomolar concentrations⁴. All four peptides were recognized by the same cytotoxic T lymphocyte line⁴⁶.

¶ The residue at position 10 is in italics. This peptide might be an exceptional A-2 restricted epitope having its C-terminal anchor residue at position 10 instead of 9.

at the very end of the cleft for Leu or Met, which have hydrophobic side chains similar to each other. For D^b we postulate a central pocket exclusively for Asn. A second pocket for Met, Ile, Leu, or Val should be at one end of the D^b-cleft. This pocket might be less specific, as some peptides reported to contain D^b-restricted epitopes do not fit to the motif at this position. Alternatively, these peptides may represent examples of a minority of epitopes having their C-terminal hydrophobic anchor residue at position 10 or 11 (Table 2c). L^d molecules should contain a pocket for Pro at one end of the groove.

The mechanism by which MHC-restricted epitopes are produced is unclear, even when they stem from known proteins.

As all epitopes appear to have hydrophobic C termini, whereas the N termini are highly variable, the hydrophobic end might be a site for an MHC-independent protease⁸, producing larger precursor peptides, whose N termini are trimmed after binding to MHC⁸.

Strictly allele-specific peptide motifs seem to be contradicted by data obtained from assays measuring MHC binding to solid-phase peptides (for example, refs 19, 20); later studies however suggest that this reflects the artificial situation in which these assays are done. We have also shown that artificial peptides (not adhering to MHC restricted motifs) binding to MHC molecules, and even inducing T-cell responses are not processed

and presented by cells expressing the respective protein²¹. From this we conclude that more peptides can bind to MHC molecules than are processed by cells. Taking into account the strong adherence of naturally processed peptides to the MHC-restricted motifs, we also conclude that these motifs are not 'peptide binding motifs', but motifs representing the outcome of processing, which is likely to include MHC-dependent and MHC-independent protease activities^{8,9} and transport mechanisms²²⁻²⁴. Intracellular binding to MHC molecules of epitopes correctly processed without participation of MHC molecules cannot entirely be excluded, although it is unlikely, given our

inability to find intracellularly correctly processed epitopes independent of MHC⁸. Thus peptide binding to MHC molecules is a necessary requirement for a peptide being an MHC-restricted epitope but is not sufficient on its own.

Knowledge of the peptide motifs of individual MHC alleles should help exact T-cell epitope predictions (K.F. *et al.*, manuscript submitted; K. Deres *et al.*, unpublished) and help with synthetic or recombinant vaccine development and potentially also for intervention in autoimmune diseases or graft rejection. □

Received 6 February; accepted 4 April 1991.

1. Zinkernagel, R. M. & Doherty, P. C. *Nature* **248**, 701-702 (1974).
2. Townsend, A. R. *et al. Cell* **44**, 959-968 (1986).
3. Bjorkman, P. J. *et al. Nature* **329**, 512-518 (1987).
4. Rötzschke, O. *et al. Nature* **348**, 252-254 (1990).
5. Van Bleek, G. M. & Nathenson, S. G. *Nature* **348**, 213-216 (1990).
6. Garrett, T. P. J., Saper, M. A., Bjorkman, P. J., Strominger, J. L. & Wiley, D. C. *Nature* **343**, 692-696 (1989).
7. Rötzschke, O., Falk, K., Wallny, H.-J., Faath, S. & Rammensee, H.-G. *Science* **249**, 283-287 (1990).
8. Falk, K., Rötzschke, O. & Rammensee, H.-G. *Nature* **348**, 248-251 (1990).
9. Shimonkevitz, R., Kappler, J., Marrack, P. & Grey, H. *J. exp. Med.* **158**, 303-316 (1983).
10. Demotz, S., Grey, H. M., Appella, E. & Sette, A. *Nature* **343**, 682-684 (1989).
11. Bjorkman, P. J. *et al. Nature* **329**, 506-512 (1987).
12. Delisi, C. & Berzofsky, J. A. *Proc. natn. Acad. Sci. U.S.A.* **82**, 7048-7052 (1985).
13. Rothbard, J. B. & Taylor, W. R. *EMBO J.* **7**, 93-100 (1988).
14. Cornette, J. L., Margalit, H., Delisi, C. & Berzofsky, J. A. *Meth. Enzym.* **178**, 611-633 (1989).
15. Sette, A. *et al. Proc. natn. Acad. Sci. U.S.A.* **86**, 3296-3300 (1989).
16. Maryanski, J. L., Verdini, A. S., Weber, P. C., Salemme, F. R. & Corradin, G. *Cell* **60**, 63-72 (1989).
17. Bastin, J., Rothbard, J., Davey, J., Jones, I. & Townsend, A. *J. exp. Med.* **165**, 1508-1523 (1987).
18. Bjorkman, P. J. & Davis, M. M. *Cold Spring Harb. Symp. quant. Biol.* **54**, 365-374 (1989).
19. Bouilliot, M. *et al. Nature* **339**, 473-475 (1989).
20. Frelinger, J. A., Gotch, F. M., Zweerink, H., Wain, E. & McMichael, A. J. *J. exp. Med.* **172**, 827-834 (1990).
21. Schild, H., Rötzschke, O., Kalbacher, H. & Rammensee, H.-G. *Science* **247**, 1587-1589 (1990).
22. Townsend, A. *et al. Nature* **340**, 443-448 (1989).
23. Elliott, T., Townsend, A. & Cerundolo, V. *Nature* **348**, 195-197 (1990).
24. Cerundolo, V. *et al. Nature* **345**, 449-452 (1990).
25. Rüsch, E., Kuon, W. & Hämmerling, G. J. *Trans. Proc.* **15**, 2093-2096 (1983).

26. Lemke, H., Hämmerling, G. J. & Hämmerling, U. *Immunol. Rev.* **47**, 175-206 (1979).
27. Ozato, K. & Sachs, D. H. *J. Immun.* **126**, 317-321 (1981).
28. Parham, P. & Brodsky, F. M. *Hum. Immun.* **3**, 277-299 (1981).
29. Taylor, P. M., Davey, J., Howland, K., Rothbard, J. B. & Askonas, B. A. *Immunogenetics* **26**, 267-272 (1987).
30. Braciale, T. J. *et al. J. exp. Med.* **166**, 678-692 (1987).
31. Braciale, T. J., Sweetser, M. T., Morrison, L. A., Kittleson, D. J. & Braciale, V. L. *Proc. natn. Acad. Sci. U.S.A.* **86**, 277-281 (1989).
32. Kuwano, K., Braciale, T. J. & Ennis, F. A. *FASEB J.* **2**, 2221 (1988).
33. Maryanski, J. L., Pala, P., Cerottini, J. C. & Corradin, G. *J. exp. Med.* **167**, 1391-1405 (1988).
34. Maryanski, J. L., Pala, P., Corradin, G., Jordan, B. R. & Cerottini, J. C. *Nature* **324**, 578-579 (1986).
35. Sibille, C. *et al. J. exp. Med.* **172**, 35-45 (1990).
36. Romero, P. *et al. Nature* **341**, 323-326 (1989).
37. Weiss, W. R. *et al. J. exp. Med.* **171**, 763-773 (1990).
38. Kast, W. M. *et al. Cell* **59**, 603-614 (1989).
39. Oldstone, M. B. A., Whitton, J. L., Lewicki, H. & Tishon, A. *J. exp. Med.* **168**, 559-570 (1988).
40. Tevethia, S. S. *et al. J. Virol.* **64**, 1192-1200 (1990).
41. Carbone, F. R. & Bevan, M. J. *J. exp. Med.* **169**, 603-612 (1989).
42. Schumacher, T. N. M. *et al. Cell* **62**, 563-567 (1990).
43. Walker, B. D. *et al. Proc. natn. Acad. Sci. U.S.A.* **86**, 9514-9518 (1989).
44. Gotch, F., McMichael, A. & Rothbard, J. *J. exp. Med.* **168**, 2045-2057 (1988).
45. Santos-Aguado, J., Crimmins, M. A. V., Mentzer, S. J., Burakoff, S. J. & Strominger, J. L. *Proc. natn. Acad. Sci. U.S.A.* **86**, 8936-8940 (1989).
46. Claverie, J. M. *et al. Eur. J. Immun.* **18**, 1547-1553 (1988).
47. Falk, K. *et al. J. exp. Med.* (in the press).

ACKNOWLEDGEMENTS. We thank J. Klein for support, A. Townsend for cell lines, S. Faath for technical assistance, and L. Yakes for typing the manuscript. This work was supported by Sonderforschungsbereich 120 and 323.

LETTERS TO NATURE

Statistical evidence for a galactic origin of gamma-ray bursts

J. L. Atteia*, C. Barat*, E. Jourdain*, M. Niel*, G. Vedrenne*, N. Blinov*, A. Chernenko†, V. Dolidze†, A. Kozlenkov†, A. Kuznetsov†, I. G. Mitrofanov†, A. Pozanenko†, R. Sunyaev† & O. Terekhov†

* Centre d'Etude Spatiale des Rayonnements, 9 Avenue du Colonel Roche, BP 4346, 31029 Toulouse Cedex, France

† Space Research Institute, Profsoyuznaya 84/32, 117810 Moscow, USSR

ASTRONOMICAL bursts of gamma-rays (GRBs) were first discovered 20 years ago, and ~100 are recorded every year by satellite-borne instruments. Bursts last for at most a few seconds, recur only on a timescale of years, if at all, and come from objects which have remained undetected at all other wavelengths. It has been impossible to establish a distance scale for GRBs, and no association with known astronomical objects has been demonstrated. Here we analyse the spatial distribution of GRBs^{1,2} with a view to understanding their true radial distribution. Our data consist of three samples, totalling 244 GRBs, obtained by three French-Soviet experiments flown on the Venera 13 and 14 and the Phobos missions. We conclude that the underlying GRB distribution is not uniformly distributed in space, but falls off with distance. Our analysis of the weak sources in particular suggests that GRBs are associated with the galactic plane.

There has recently been interest in using the $V/V_{\max}$ distribution to study the radial distribution of the sources^{3,4}, where for a given burst V is the volume of the smallest sphere containing the source, and $V_{\max}$ is the maximum volume accessible to the

instrument. This ratio is defined as follows: if D is the (unknown) distance of the GRB source and $D_{\max}$ the maximum distance at which it can be detected, then $V/V_{\max} = (D/D_{\max})^3$, and as the number of detected counts, C , scales as D^{-2} , we can write $V/V_{\max} = (C_{\max}/C_{\min})^{-3/2}$ where $C_{\max}$ is the maximum number of counts and $C_{\min}$ the minimum number required to detect a burst. The $V/V_{\max}$ distribution is free of instrumental selection effects as long as $C_{\max}$ and $C_{\min}$ are calculated in the same time intervals and energy range as those used to trigger the experiment³ (detailed analysis shows that some bias can still affect the $V/V_{\max}$ distribution; Hartmann *et al.*, manuscript in preparation). If the number of sources per unit volume is constant, corresponding to an infinite uniform density population, the $V/V_{\max}$ values are uniformly distributed between 0 and 1. Here we present the $V/V_{\max}$ distribution for bursts detected by the Signe experiments on the Soviet Venera 13 and 14 probes, and by the Lilas and Apex experiments on the Phobos probe. We use both the mean value $\langle V/V_{\max} \rangle$ and a Kolmogorov-Smirnov (KS) test to compare the observed $V/V_{\max}$ distribution with a uniform one. Two of the three data sets exhibit a significant deviation from a uniform distribution, which may be interpreted as a deficit of sources at large distances. We show that the faintest sources (in the $V/V_{\max}$ sense) seem to be concentrated in the galactic plane, evidence which supports a galactic-disk distribution of GRB sources. The $V/V_{\max}$ analysis has been used previously to study the radial distribution of burst samples⁴⁻⁷, but here we associate a low $\langle V/V_{\max} \rangle$ value with the observation of angular anisotropy.

Table 1 gives the main characteristics of the three experiments used here (details available for Signe⁸ and in preparation for Lilas and Apex (C. Barat *et al.*, manuscript in preparation)). The Signe experiment had four identical detectors, the Lilas

TABLE 1 Principal characteristics of the three samples of bursts

Experiment	Venera-Signe	Apex	Lilas
Detector	Nal (cyl.)	CsI (cyl.)	Nal (cyl.)
Size: diameter $\times$ height (cm)	9 $\times$ 3.7	10 $\times$ 10	5.3 $\times$ 3
Energy range	50–350 keV	120–700 keV	20–300 keV
Lifetime	11/81–04/83	08/88–03/89	08/88–03/89
Number of triggers	1,930	509	616
Number of bursts	169	58	47
Burst detection frequency (yr ⁻¹)	~125	~115	~90

experiment two. When $V/V_{\max}$ analysis is performed on a burst sample containing events recorded by several detectors, it must be done as follows: the lowest $V/V_{\max}$ value is taken if, as here, the sample contains all events seen by one or more detectors; if the sample is restricted to events detected by all the detectors, then the highest $V/V_{\max}$ value is taken. The trigger algorithm was the same for the three experiments: a trigger occurs when the number of counts in Δt seconds in an energy range ΔE exceeds the background level (measured over 1 or 2 minutes) by more than N (usually 7 or 8) standard deviations. Usually several values of Δt are used simultaneously, but for consistency we have restricted the present analysis to bursts capable of triggering the instruments in a 1-s window common to all three instruments. Table 1 indicates that most triggers are not due to cosmic bursts. As it is important to retain all the true GRBs and nothing else, we have paid particular attention to selection of events. For the Signe experiment, 'false triggers' are caused either by heavy particles or solar events. The former can be identified by their characteristic exponential decay and soft spectrum; solar bursts are readily eliminated by comparison with data on solar activity. After this primary event selection, comparison with GRB data from other instruments showed that 73% of Signe GRBs were detected by two spacecraft at least. Because of its high energy threshold (see Table 1), Apex detected only 10 solar events, all associated with strong solar flares and easily recognizable; most of the event triggers were identified as being due to heavy particles traversing the detector and were easily eliminated (as for Signe). The Lilas experiment was found to trigger only on solar events and cosmic bursts. As the two Lilas detectors had the same orientation with respect to the sun, the solar-flare-induced events were characterized by a similar flux on both detectors and by their soft spectrum. We used a procedure based on these observational attributes to eliminate solar events. The efficiency of this selection was checked by

comparison with data on solar activity from the World Data Center (the region of the Sun visible from Phobos was identical to that visible from the Earth during 75% of the mission). Only four events of the 616 recorded by Lilas had their classification changed by this procedure. We are therefore confident that more than 90% of our lists of bursts from the three experiments are true GRBs.

Figure 1 shows $V/V_{\max}$ cumulative distribution for the entire set of 244 bursts together with the theoretical distribution for a uniform population of sources. If the spatial distribution of the sources is uniform, the $V/V_{\max}$ distribution is also uniform. If the spatial distribution is not uniform, the situation is much more complex, and the $V/V_{\max}$ distribution depends on factors such as the spatial distribution of the sources or their luminosity function. Here we use the data simply to check the uniformity of the spatial distribution of GRB sources; more information can be extracted from the shape of the observed curve if detailed simulations of the GRB population, beyond the scope of this paper, are carried out. Figure 1 shows that the data are not compatible with a uniform distribution of GRBs. The hypothesis of spatial uniformity is rejected with a confidence level $\geq 4\sigma$ as indicated by a KS test (probability = 5×10^{-6}) and by the value $\langle V/V_{\max} \rangle = 0.4 \pm 0.018$; the uncertainty is $(12N)^{-1/2}$, where N is the number of bursts in the sample³. The curve indicates a deficit of sources at large distances, as only one-third of the sources have $V/V_{\max} > 0.5$. Figure 2, which presents the cumulative $C_{\max}/C_{\min}$ distribution for bursts recorded with Signe⁷, shows the same effect: the region with $C_{\max}/C_{\min} > 2$, with ~80 events, is characterized by a slope of -1.5, corresponding to uniform population of sources, whereas for $C_{\max}/C_{\min} < 2$, the slope decreases, thus indicating the deficit (significance 3.2σ).

These observations could be explained by sources at cosmological distances⁹, by a galactic halo with a decreasing density of sources or by a galactic disk population. One way to distinguish between these possibilities is to use the determination of source positions on the sky, but previous angular distribution analysis^{10,11} did not reveal any galactic structure. Of the 169 GRBs detected by Venera 13 and 14, 51 events have been localized¹². We have divided these events into two classes of almost equal significance (Fig. 2): 24 'bright bursts' with $C_{\max}/C_{\min} \geq 2.5$ and 27 'faint bursts' with $C_{\max}/C_{\min} < 2.5$ (the boundary between the classes is close to the point where the curve $C_{\max}/C_{\min}$ flattens). The main idea of this subdivision is to remove brighter and *a priori* closer sources which may blur possible structures (a similar approach is used in ref. 13). We have then used the method of Hartmann and Epstein¹¹ to study the angular distribution of sources within each class. The spatial distribution is described by its first two moments and more

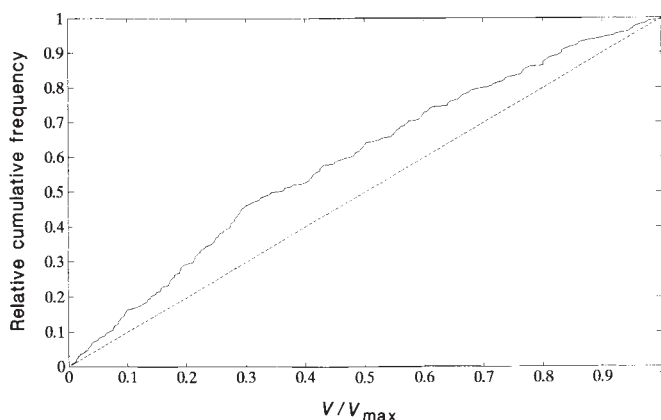


FIG. 1 $V/V_{\max}$ distribution of 244 gamma-ray bursts from all three detectors. Note the deficit of sources at large distances with respect to a uniform population (only one-third of the sources have $V/V_{\max} > 0.5$). The mean value, $\langle V/V_{\max} \rangle$, is 0.419 ± 0.018 .

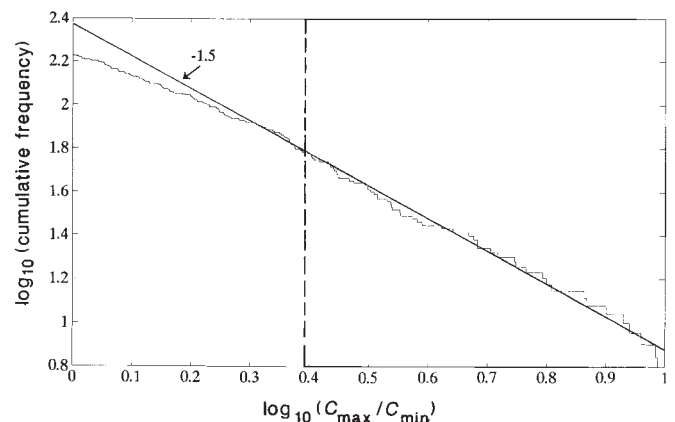


FIG. 2 Distribution of $C_{\max}/C_{\min}$, for 169 gamma-ray bursts detected by the Signe instruments on Venera 13 and 14. A deficit of weak bursts is apparent by comparison with the $-3/2$ slope expected for a uniform population. The dashed line separates 'bright' from 'faint' bursts (see text).

precisely by three parameters: P , the magnitude of the dipole vector; η , the absolute value of the difference between the two most similar eigenvalues of the quadrupole tensor; and ζ , the remaining eigenvalue of the quadrupole tensor (see ref. 11 for details). No strong anisotropy is present for the class of 'bright bursts' ($P = 0.07$, $\eta = 0.258$, $\zeta = 0.328$). The distribution of 'faint bursts' ($\zeta = -0.636$, $P = 0.146$, $\eta = 0.096$) is much more interesting. The high absolute value of ζ and the low absolute value of η indicate respectively that faint sources are concentrated in a plane and have good axial symmetry. A Monte Carlo simulation shows that when 27 bursts are randomly distributed on the celestial sphere, such a concentration in a plane is found in only eight cases out of 10^4 . Furthermore, the normal to this plane is $9 \pm 10^\circ$ from the galactic pole. The events studied here were detected by both Signe and Konus instruments on Venera and, as the Signe detectors lie in the ecliptic plane and have a 1.5-yr lifetime, we do not believe the experiment to be biased in favour of the galactic plane. We therefore interpret this result as showing that GRBs have a galactic origin and that the departure of the $V/V_{\max}$ curves from uniformity is due to the decreasing density of sources for increasing distances from the galactic plane. The sampling depth of current instruments is a few times the scale height of the distribution deduced from these results. Moreover, the absence of centre-anticentre anisotropy for faint sources (indicated by low η) indicates that the sources are no further than a few kiloparsecs from the Sun. Finally, the difference in spatial distribution between 'bright' and 'faint' sources proves that the mean $V/V_{\max}$ value of a sample of bursts provides a fair estimate of its relative distance (obviously this is not true for individual bursts).

Models that explain GRBs as due to the interaction of old neutron stars with the interstellar medium^{14,15} predict a scale height of the sources of ~ 200 pc. This height is consistent with the distribution of the interstellar gas and of the low-velocity neutron stars which are capable of accreting at sufficiently high rates¹. As the sampling depth of current observations is a few times the scale height, the GRB sources are visible within several hundred parsecs of the Sun, and two observational results should be carefully considered. First, the absence of angular clustering of localized sources¹⁰ does not seem compatible with the known inhomogeneity of the local interstellar medium. Second, the absence of persistent X-ray counterparts¹⁶, if confirmed by ROSAT, may be a problem for these models.

These last considerations show the importance of measuring the spatial distribution of GRBs if we are to understand these objects. We expect the BATSE experiment¹⁷, which has high sensitivity and localization capability, on the Gamma Ray Observatory to provide further information, and the concentration of sources towards the galactic plane reported here should become apparent after several months of operation. $\square$

Received 30 October 1990; accepted 2 April 1991.

1. Paczynski, B. *Astrophys. J.* **348**, 485–494 (1990).
2. Hartmann, D., Epstein, R. I. & Woosley, S. E. *Astrophys. J.* **348**, 625–633 (1990).
3. Schmidt, M., Higdon, J. C. & Hueter, G. *Astrophys. J.* **329**, L85–L87 (1988).
4. Higdon, J. C. & Schmidt, M. *Astrophys. J.* **355**, L13–L17 (1990).
5. Higdon, J., Matz, S. M., Share, G. H., Messina, D. & Iadiccio, A. *Proc. TAOS Workshop on Gamma-Ray Bursts* (eds Cheng, H., Epstein, R. I. & Feniman, E. E.) (Cambridge University Press, UK, 1990).
6. Mitrofanov, I. et al. *Proc. TAOS Workshop on Gamma-Ray Bursts*, (eds. Cheng, H., Epstein, R. I. & Feniman, E. E.) (Cambridge University Press, UK, 1990).
7. Mitrofanov, I. et al. *Soviet Astron. J.* **68**, (1991).
8. Barat, C. et al. *Astrophys. Space Sci.* **75**, 83–91 (1981).
9. Paczynski, B. *Astrophys. J.* **306**, L43–L46 (1986).
10. Hartmann, D. & Blumenthal, G. *Astrophys. J.* **342**, 521–526 (1989).
11. Hartmann, D. & Epstein, R. I. *Astrophys. J.* **346**, 960–966 (1989).
12. Golenetskii, S. V. et al. preprint no. 1026, A. F. Ioffe Phys.-Tech. Inst., Leningrad (1986).
13. Pozanenko, A. & Mitrofanov, I. *Astrophys. Space Sci.* **164**, 277–283 (1990).
14. Hameury, J. M., Bonazzola, S., Heyvaerts, J. & Ventura, J. *Astr. Astrophys.* **111**, 242–251 (1982).
15. Blaes, O., Blandford, R., Madau, P. & Koonin, S. *Astrophys. J.* **363**, 612–627 (1990).
16. Pizzichini, G. et al. *Astrophys. J.* **301**, 641–649 (1986).
17. Fishman, G. J. et al. *Proc. Gamma Ray Observatory Science Workshop* (Goddard Space Flight Center, Greenbelt, Maryland, 1989).

ACKNOWLEDGEMENTS. We thank J.-M. Hameury, S. Bonazzola, K. Hurley and D. Hartmann for helpful discussions, and A. J. Dean for reading the manuscript.

Formation of hierarchical multiple protostellar cores

Alan P. Boss

Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road, NW, Washington, DC 20015, USA

BINARY pre-main-sequence stars^{1,2} seem to occur as frequently as binary main-sequence stars^{3,4}; triple pre-main-sequence star systems have also been detected⁵, and hierarchical main-sequence multiple stars continue to be identified⁶. (Hierarchical systems contain both closely spaced stars and stars orbiting at much greater distances.) These observations suggest that essentially all binary stars were formed before the main-sequence phase of evolution. The detection of a number of binary young stellar objects^{7,8} seems to indicate that binary formation must occur no later than the protostellar phase (further observations are needed to establish if this is the case for multiple star systems). Here I describe numerical hydrodynamical calculations showing that stable hierarchical systems of multiple protostellar cores can form through gravitationally driven fragmentation during the collapse of an isolated gas cloud, suggesting that the hierarchical systems observed may be the result of the hydrodynamical collapse of rapidly rotating clouds.

Fragmentation during the isothermal collapse phase of protostellar evolution is one of the leading mechanisms for explaining the formation of binary and multiple stars^{9–13}. Although it is well known that fragmentation can form binary protostellar cores or trapezium-like (unstable) multiple protostellar cores^{13,14}, until now fragmentation calculations have not demonstrated the formation of a stable, hierarchical protostellar core system.

The calculations were carried out with a new explicit, Eulerian, three-dimensional hydrodynamics code which is accurate to second order in spatial finite differences (A.P.B., in preparation). To obtain the hydrodynamical fluxes, I use consistent advection and an extension in spherical coordinates of van Leer monotonic interpolation¹⁵. The code is accurate to first order in time; this should be adequate given the large number ($\sim 6 \times 10^4$) of time steps used in the models. The gravitational potential is obtained by a spherical harmonic expansion with terms up to and including $l, m = 16$. The spatial resolution is 51 grid points in radius, effectively 45 in colatitude (assuming symmetry through the equatorial plane) and 64 in azimuth (with no symmetry about the rotation axis).

The hydrodynamical equations are solved in conservation-law form, ensuring global conservation of mass and angular momentum. The code is designed to conserve angular momentum locally, as measured by the preservation of the initial angular momentum spectrum during axisymmetric collapse¹⁶. The radial momentum fluxes are corrected for the effects of volume centring in spherical geometry¹⁷. The Q^r and $Q^{\theta\theta}$ components of the artificial viscosity tensor are used to stabilize the scheme against shocks¹⁸; the artificial viscosity vanishes during homologous collapse. Convergence testing (L. S. Finn and J. F. Hawley, personal communication, 1989) on spherical and three-dimensional test cases has demonstrated the second-order accuracy of the code. The code can maintain a stable polytropic equilibrium with non-axisymmetric 'noise' of $\leq 1\%$.

The initial conditions for protostellar collapse are taken from observations of molecular cloud complexes. Dense cloud cores¹⁹ seem typical of one mode of low-mass star formation. Line widths are consistent with near-virial equilibrium²⁰, and the cores have sizes ~ 0.1 pc, mean densities $< 2 \times 10^5 \text{ cm}^{-3}$, temperatures ~ 10 K and masses $> 0.5 M_\odot$ ²¹. The cloud cores are centrally condensed and in some cases have considerable rotation²². Dense cloud cores of solar mass seem to be less affected by magnetic fields than more massive and diffuse clouds^{23,24}.

I have made calculations for a sequence of four hydrodynamical models, starting from conditions appropriate for rapidly-rotating, non-magnetic, dense cloud cores. The models are assumed to remain isothermal at the initial cloud temperature of 10 K. Each cloud has a mass of $1.0 M_{\odot}$ and an initial radial density profile $\rho_i \propto \exp(-(r/r_1)^2)$, where r_1 is a constant, chosen to make the central density 20 times as high as the boundary density. The density was given a 10% azimuthal ($\cos 2\phi$) perturbation, an initial asymmetry considerably less than that exhibited by many dense cloud cores²¹. The four models differ in their initial sizes, densities and values of $\alpha_i = E_{\text{thermal}}/|E_{\text{gravitational}}|$: model C1 has a radius $R = 0.024$ pc and $\alpha_i = 0.40$, model C4 has $R = 0.016$ pc, $\alpha_i = 0.26$, model C7 has $R = 0.0081$ pc, $\alpha_i = 0.13$ and model C10 has $R = 0.0042$ pc, $\alpha_i = 0.060$. In each model, the cloud initially rotates as a solid body with an angular velocity sufficient to yield $\beta_i = E_{\text{rotational}}/|E_{\text{gravitational}}| = 0.16$.

Models C1 and C4 are initially the closest to virial equilibrium ($\alpha + \beta = 1/2$) and could represent clouds that begin to collapse after ambipolar diffusion has lessened the magnetic field support; models C7 and C10 may better represent clouds that are

induced to collapse by rapid compression (for example, by shock waves).

Models C7 and C10 collapsed and formed intermediate bar configurations before fragmenting into binary protostellar cores, as in previous fragmentation studies²⁵. Models C1 and C4 behaved quite differently: they collapsed to form an intermediate disk which subsequently began to fragment. Figure 1 shows the evolution of model C4. Binary fragmentation followed by sub-fragmentation of these binaries is evident. The four protostellar cores produced in model C4 have masses in the range 0.005 – $0.01 M_{\odot}$, $\alpha \approx 0.1$ – 0.4 and $\beta \approx 0.01$ – 0.1 .

The different results of models C1 and 4, and C7 and 10 must be caused by the differing initial conditions. Models C7 and C10 started with considerably lower thermal pressure than C1 and C4, and as a consequence the initial bar-like ($m = 2$) perturbation was able to grow rapidly and produce a narrow spindle. Models C1 and C4 started much closer to virial equilibrium, and as a consequence remained fairly axisymmetric until a rotationally flattened disk began to form. Higher-order modes ($m = 4$ and above) could then participate in the sub-fragmentation of model C4. Model C1 was not computed quite as far as

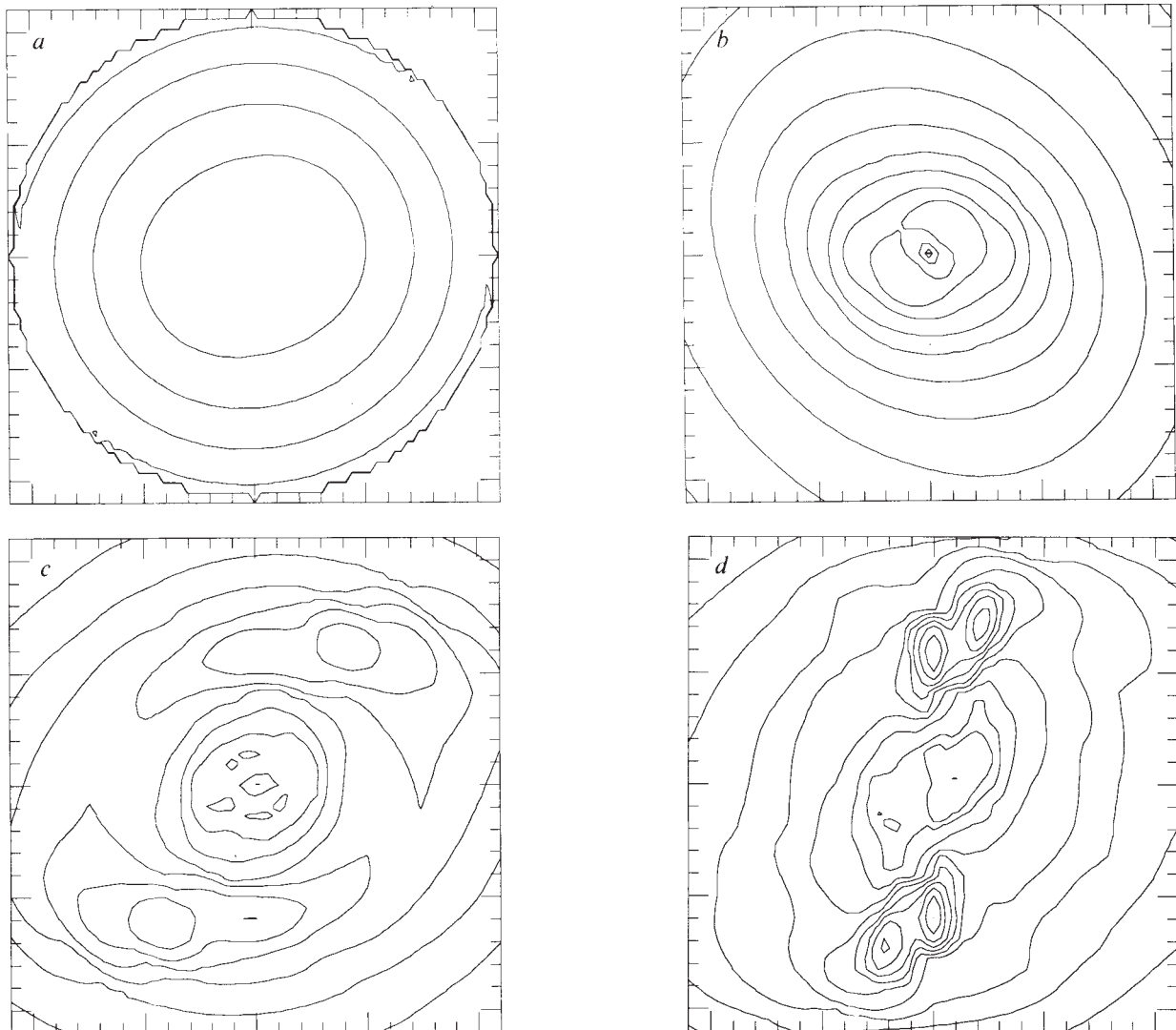


FIG. 1 Density contours in the equatorial plane at four different times for model C4: *a*, $0.345 t_{\text{ff}}$, *b*, $1.379 t_{\text{ff}}$, *c*, $1.421 t_{\text{ff}}$, *d*, $1.443 t_{\text{ff}}$, where t_{ff} = free fall time = 1.6×10^4 yr. The logarithm of the density ρ is plotted; the density changes by a factor of two between contours. Density maxima are seen in *c* and *d*, reflecting the formation of hierarchical multiple protostellar cores (density minima are marked with minus signs). For *a*, *b*, *c* and *d*, respectively,

$\log_{10} \rho_{\text{max}} = -16.7, -13.5, -12.4$ and -11.5 (ρ_{max} in g cm^{-3}), and the radius of the region plotted is 5×10^{16} , 0.70×10^{16} , 0.20×10^{16} and 0.20×10^{16} cm. Counterclockwise rotation is assumed. The box side subtends 101 radial grid points for *a*, 15 for *b*, *c* and *d*; the jagged cloud boundary in *a* is an artefact of the plotting routine.

model C4 but seemed likely to experience a similar sub-fragmentation. Paradoxically, then, clouds starting close to virial equilibrium may produce more fragments than clouds starting well below virial equilibrium ($\alpha + \beta < 1/2$), at least during these early phases.

The four models all started with fairly rapid rotation ($\beta_i = 0.16$), although rotation is not commonly detected in dense molecular clouds. Very slowly rotating, centrally-condensed clouds ($\beta_i \ll 0.1$) are unlikely to fragment²⁵ unless the initial thermal energy is very small. If hierarchical protostellar systems are formed by rotation-driven fragmentation, as demonstrated here, the fraction of dense molecular clouds with appreciable rotation ($\beta_i \approx 0.01$ – 0.1) must be related to the fraction of protostars in such systems. Evidently a sensitive survey of molecular cloud rotation rates and further searches for multiple protostellar systems are needed to decide if rotation-driven fragmentation makes an important contribution to the formation of multiple systems.

At the time of Fig. 1d, the fragment masses in model C4 are appropriate for newly formed protostellar cores; subsequent evolution will presumably result in the accretion of the bulk of the protostar masses from the dense cloud core. It is possible that the protostellar cores will merge during this subsequent period as a result of effects such as gas accretion and orbital decay through gravitational torques from the cloud envelope. I do not predict subsequent fragment mergers for model C4, however, largely because the fragmentation occurred primarily during the isothermal phase. Protostellar mergers involve a race between collapse of the protostellar cores to stellar sizes and orbital decay through gravitational torques, and isothermal protostars seem to collapse too rapidly to merge²⁶. Gas drag and accretion of gas (generally with higher angular momentum) are not thought to modify this conclusion appreciably²⁶. Recent simulations²⁷ of binary orbital decay through gaseous disk interactions predicted orbital decay times much longer than those used to reach this conclusion²⁶. Other recent simulations have shown that protostellar gas disks can be dispersed if nearby protostellar cores fly past at high velocity²⁸. In contrast, quasi-equilibrium (nonisothermal) protostars may undergo orbital mergers, depending on the details of the model²⁵. Although the fragments in model C4 seem likely to survive, we cannot definitely conclude that a hierarchical stellar system is formed by fragmentation until it is shown that these cores do survive to become stars.

Model C4 seems to be the first direct calculation of hierarchical fragmentation²⁹ leading to formation of a stable hierarchical system of multiple protostellar cores. The prediction of stability for the hierarchical system comes from the orbital spacings evident in Fig. 1d: the ratio of the long-period binary separation (a_2) to that of the short-period binary (a_1) is ~ 5.5 . This ratio exceeds the critical value for orbital stability of $a_2/a_1 \approx 4.6$ obtained from the three-body stability criterion of Graziani and Black^{30,31} and the fragment masses calculated in the model; observed hierarchical systems have $a_2/a_1 > 5.0$ – 5.5 . $\square$

15. Norman, M. L. & Winkler, K.-H. in *Astrophysical Radiation Hydrodynamics* (eds Winkler, K.-H. & Norman, M. L.) 187–221 (Reidel, Dordrecht, 1986).
16. Norman, M. L., Wilson, J. R. & Barton, R. T. *Astrophys. J.* **239**, 968–981 (1980).
17. Mönchmeyer, R. & Müller, E. *Astr. Astrophys.* **217**, 351–367 (1989).
18. Tscharnuter, W. M. & Winkler, K.-H. *Comput. Phys. Comm.* **18**, 171 (1979).
19. Myers, P. C. & Benson, P. J. *Astrophys. J.* **266**, 309–320 (1983).
20. Myers, P. C. *Astrophys. J.* **270**, 105–118 (1983).
21. Benson, P. J. & Myers, P. C. *Astrophys. J. Suppl. Ser.* **71**, 89–108 (1989).
22. Arquilla, R. & Goldsmith, P. F. *Astrophys. J.* **303**, 356–374 (1986).
23. Heyer, M. H. *Astrophys. J.* **324**, 311–320 (1988).
24. Myers, P. C. & Goodman, A. A. *Astrophys. J.* **329**, 392–405 (1988).
25. Boss, A. P. *Astrophys. J.* **319**, 149–161 (1987).
26. Boss, A. P. *Mon. Not. R. astr. Soc.* **209**, 543–567 (1984).
27. Artymowicz, P., Clarke, C. J., Lubow, S. H. & Pringle, J. E. *Astrophys. J.* (in the press).
28. Clarke, C. J. & Pringle, J. E. *Mon. Not. R. astr. Soc.* (in the press).
29. Bodenheimer, P. *Astrophys. J.* **224**, 488–496 (1978).
30. Graziani, F. & Black, D. C. *Astrophys. J.* **251**, 337–341 (1981).
31. Black, D. C. *Astr. J.* **87**, 1333–1337 (1982).

Use of a neural network to control an adaptive optics system for an astronomical telescope

D. G. Sandler, T. K. Barrett, D. A. Palmer,
R. Q. Fugate* & W. J. Wild*†

Thermo Electron Technologies, 9550 Distribution Avenue, San Diego, California 92121, USA

* Starfire Optical Range, Air Force Phillips Laboratory, Albuquerque, New Mexico 87117, USA

ANGEL *et al.*¹ recently showed how an artificial neural network could be used to measure optical phase distortion induced by atmospheric turbulence, and demonstrated by numerical simulation that such a system could be used to control the six 1.8-m mirrors of the Multiple Mirror Telescope by constantly adjusting them to compensate for atmospheric distortion of the image. The neural network estimates the phase distortion using two images of a reference star, or of a laser-produced guide star², one image being at the best focus of the telescope while the other is intentionally out of focus. Here we report the successful test of a neural network with a real star. We applied a neural network to in- and out-of-focus images of Vega obtained with the 1.5-m single-mirror telescope at the Starfire Optical Range of the Air Force Phillips Laboratory near Albuquerque, New Mexico. The experimental results agree well with phase reconstructions obtained simultaneously with a conventional wave-front sensor.

The neural network has several advantages over conventional Hartmann sensors or shearing interferometers³. It operates directly on the optical point-spread function (PSF) of the system, which is the quantity of most interest. The cameras used to record the double-image data can be located near the telescope aperture and the astronomical viewing camera, minimizing uncommon (that is, non-shared) optics which lead to additional optical distortion and photon losses. Thus the neural network, in conjunction with an active deformable mirror⁴ to compensate for atmospheric distortion, is a simple, inexpensive way to implement adaptive optics on astronomical telescopes in the near term.

A neural network is a parallel computational architecture composed of individual elements, or 'neurons', modelled loosely on biological neurons^{5–7}. A variety of neural network architectures exists; each is specified by the interconnections between the processing nodes, the operations performed by each node and the method of 'training' or optimizing the network performance. Our phase-recovery network employs multiple perceptrons⁸. Its architecture is shown schematically in Fig. 1. The network is trained by adjusting the connection weights using the back-propagation algorithm⁹, which repeatedly presents the

Received 21 January; accepted 25 March 1991.

1. Mathieu, R. D., Walter, F. M. & Myers, P. C. *Astr. J.* **98**, 987–1001 (1989).
2. Chen, W. P., Simon, M., Longmore, A. J., Howell, R. R. & Benson, J. A. *Astrophys. J.* **357**, 224–230 (1990).
3. Abt, H. A. *Ann. Rev. astr. Astrophys.* **21**, 343–372 (1983).
4. Duquenois, A. & Mayor, M. in *Proc. XI Eur. Regional Astr. Meeting IAU* (ed. Vazquez, M.) (Cambridge University Press, UK, in the press).
5. Zinnecker, H. in *Low Mass Star Formation & Pre-Main-Sequence Objects* (ed. Reipurth, B.) 447–469 (European Southern Observatory, Garching, 1989).
6. McCarthy, D. W. Jr., Henry, T. J., McLeod, B. & Christou, J. C. *Astr. J.* **101**, 214–219 (1991).
7. Wootten, A. *Astrophys. J.* **337**, 858–864 (1989).
8. Sasselov, D. D. & Rucinski, S. M. *Astrophys. J.* **351**, 578–582 (1990).
9. Boss, A. P. *Comments Astrophys.* **12**, 169–190 (1988).
10. Boss, A. P. in *Close Binaries* (eds Sahade, J., McCluskey, G. & Kondo, Y.) (Kluwer, Dordrecht, in the press).
11. Pringle, J. E. *Mon. Not. R. astr. Soc.* **239**, 361–370 (1989).
12. Pringle, J. E. in *NASA ASI Conf. Proc. Star Formation* (ed. Kylafis, N.) (Kluwer, Dordrecht, in the press).
13. Bodenheimer, P., Tohline, J. E. & Black, D. C. *Astrophys. J.* **242**, 209–218 (1980).
14. Miyama, S. M., Hayashi, C. & Narita, S. *Astrophys. J.* **279**, 621–632 (1984).

† Present Address: Department of Astronomy and Astrophysics, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA.

network with training data and minimizes the squared error between the known solution and the network output, using a gradient search technique.

Several considerations affected our choice of neural network. First, phase recovery from stellar image data is a very nonlinear problem, because the PSF is the squared modulus of the Fourier transform of $e^{i\phi}$, where ϕ is the phase distortion across the telescope aperture. Note from Fig. 1 that sigmoid functions, $(1 + e^{-x})^{-1}$, where x represent the weighted inputs into a given

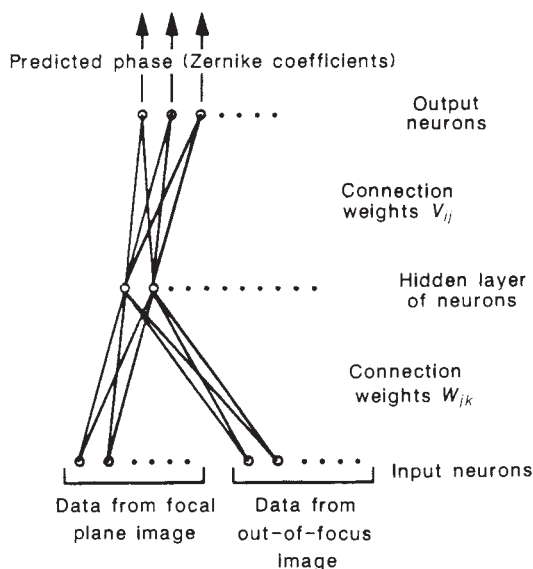


FIG. 1 Schematic representation of the operation of the artificial neural network used to recover phase. The first layer of neurons takes the intensity (camera data) from the two focal planes and passes these values to the second layer. Each of the k th input nodes is connected to the j th node in the second or hidden layer by a connection weight W_{jk} which multiplies the input intensity I_k . Each neuron in the hidden layer performs the sum $S_j = \sum_k W_{jk} I_k$ and passes the result through a nonlinear sigmoid function f . The result of each hidden node, $O_j = f(S_j)$, is then multiplied by another connection weight V_{ij} and passed to the i th output node. The sum, $Z_i = \sum_j V_{ij} O_j$, is the network's estimate for the i th Zernike mode.

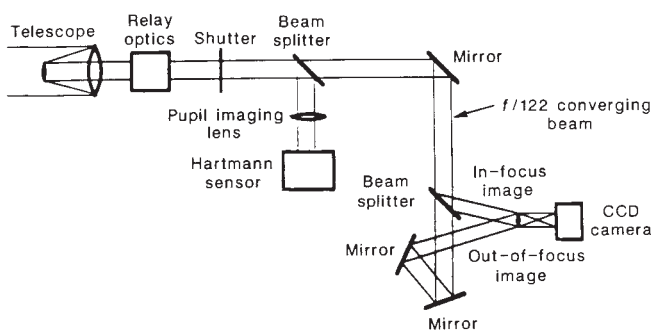


FIG. 2 Optical beam train used to collect the stellar data. Starlight enters the beam train through the 1.5-m telescope and passes through a coude path and a relay optical system whose output is an $f/122$ converging beam. A beam splitter directs a portion of the starlight through a pupil imaging system to the Hartmann sensor. The remaining light is directed to a second beam splitter which reflects a portion of the light directly to the CCD camera, creating the in-focus stellar image. The light that passes through the second beam splitter reflects off two steering mirrors and onto the same camera, creating an out-of-focus image displaced from the in-focus image. The difference in path length between the two images induces the defocus, 1.3 waves r.m.s., needed for the operation of the neural network.

neuron, are used in the 'hidden' or middle layer of the network. They mimic the chemical synapses of the brain by setting a threshold on the output of each neuron (that is, the response of the neuron saturates, or becomes constant, above a certain input level). This neuronal transfer function enables the network to learn nonlinear mappings. Second, supervised training algorithms, such as back-propagation, have so far proved superior to unsupervised or self-organizing networks in learning complex functional relationships. Finally, after training, each hidden neuron corresponds to a feature of the input intensity pattern. As the algorithm is used to adjust its weights during training, the network is free to develop the features or template functions that it determines to be optimal for estimating phase aberration. These features are then correlated by the output neurons to predict the phase. This process is analogous to an optical scientist's learned ability to recognize certain optical aberrations from the shapes and features in PSF data, but this human ability is qualitative at best, and fails when the phase spectrum contains many aberrations which diffract collectively to form the image.

For our experimental tests, we chose to describe optical phase distortion in terms of orthogonal Zernike polynomials¹⁰, a set of orthogonal polynomials defined on the unit circle. They consist of products of angular functions and radial polynomials, and are directly related to common aberrations described by opticians (focus, astigmatism, coma, spherical aberration and so on). This method of describing the phase makes economical use of output neurons. The Zernike power spectrum of atmospheric turbulence falls quickly with increasing order¹¹, so that detection of only low-order (≤ 10) Zernike coefficients is sufficient to determine the phase with a small residual error.

Figure 2 is a schematic of the optical beam train that obtained the data needed to test the neural network with real stellar images. A 15-Hz closed-loop bandwidth stellar tracker was operated while recording data. As shown in Fig. 2, a Hartmann lenslet array was used to measure phase distortion across the telescope, thus providing conventional phase estimates with which to compare the neural network estimates. The atmospheric distortion was reconstructed from Hartmann sensor data at 149 points within the 1.5-m telescope aperture.

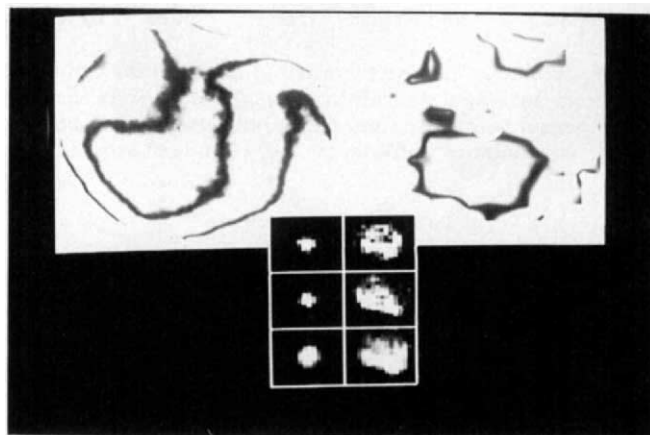


FIG. 3 An example of neural network performance showing three pairs of in- and out-of-focus far-field patterns, and two interferograms (phase contours produced by a heterodyne interferometer). The lowest pair of images are the actual camera data obtained at Starfire optical range and are the input to the neural network. The middle set of images is produced by numerically simulating the in- and out-of-focus stellar images using the phase predicted by the neural net. The top set of images is produced by numerically simulating camera data corresponding to the phase reconstructed from the Hartmann sensor. The interferogram on the left represents the phase predicted by the neural net, that on the right represents the phase measured by the Hartmann sensor.

To maximize the signal-to-noise ratio of the stellar data, pixels significantly larger than the diffraction-limited angular resolution of the telescope were used. Relatively coarse resolution suffices for determining the low-order Zernike modes from stellar images. In our case $2.44\lambda/D = 0.285$ arcsec is the diffraction-limited resolution of the Starfire Optical Range telescope at $\lambda = 0.85 \mu\text{m}$. The double images of Vega used by the neural network were 16×16 arrays of 0.412 -arcsec pixels, digitized by a single CCD (charge-coupled device) camera, with an exposure time of 2 ms.

The strength of optical turbulence is often characterized by the ratio D/r_0 , where D is the telescope diameter and r_0 is the transverse scale size below which turbulence does not degrade image resolution¹². The parameter r_0 was measured along the stellar observing path and ranged from 10 to 30 cm at $\lambda = 0.85 \mu\text{m}$, corresponding to $5 \leq D/r_0 \leq 15$. The mean value of D/r_0 during the data acquisition was 7.1.

The digitized stellar images were passed through a neural network that had been trained on simulated image data, and the network's estimate of the phase was compared with the phase reconstructed from the Hartmann sensor data. Figure 3 illustrates the network performance. Similarities are quite evident between the real image data and the far-field patterns predicted by the neural network, and between the interferograms (or contours) of the phase estimated by the neural network and by the Hartmann sensor. This figure shows qualitatively that the neural network is accurately predicting the phase at the telescope aperture.

Quantitative comparisons between the Hartmann sensor and the neural network were obtained for 107 independent realizations of atmospheric turbulence. Figure 4 shows, for these 107 samples, the phase variance per Zernike mode as reconstructed by the Hartmann sensor and the mean-square difference between the reconstructed phase and the neural network estimates of phase for modes 4 (focus) to 11 (spherical aberration). The phase-variance spectrum of the experimental Hartmann reconstructions resembles the expected Kolmogorov distribution¹¹, for $D/r_0 \approx 7$. The root-mean square difference between the network's predictions and the Hartmann-sensor reconstructions for modes 4–7 is $\lambda/14$. These data indicate that if used in an adaptive optics system, the network would reduce the mean-square wavefront error from 1.77 rad^2 to 0.78 rad^2 at most, corresponding to an increase in image resolution^{1,12} by a factor of ~ 3 .

For Zernike modes ≥ 8 there is less power in the turbulence spectrum, making it difficult to compare results from the Hartmann sensor with those from the neural network. The measurement uncertainties introduced by Hartmann-sensor noise,

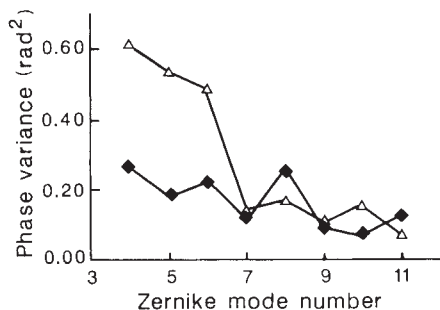


FIG. 4 Statistics for 107 independent measurements of atmospheric turbulence. Δ , variance about mean per Zernike mode for phase reconstructed from Hartmann sensor. $\blacklozenge$, $E\{[(\phi_{NN} - \phi_{NN}) - (\phi_{HS} - \phi_{HS})]^2\}$ per Zernike mode, where $E[\]$ denotes the estimate of the expected value, ϕ_{NN} is the mean phase determined by the neural network and ϕ_{HS} is the mean phase predicted by the reconstructions. Mode 4 is focus, 5 is 45° astigmatism, 6 is 0° astigmatism, 7 is x-coma, 8 is y-coma and 11 is third-order spherical.

unshared optical paths, alignment errors and aberrations in the static beam train are of the same order as the actual phase distortions introduced by the atmosphere. We note, however, that the neural network phase predictions for the higher-order modes did not diverge to unreasonable values, but remained at the right quantitative level.

We have demonstrated that a simple optical sensor with a neural network processor can measure low-order distortions created by atmospheric turbulence. Candidates for astronomical imaging using this approach include the Galactic Centre, using a foreground reference star for network inputs, and globular clusters, using the centre of the cluster or a nearby guide star. To treat objects with no suitable stellar reference, the data for the neural network can be supplied by a laser guide star, an artificial star created high in the atmosphere by laser excitation of the mesospheric sodium layer. This application of the neural network would particularly benefit infrared astronomy and speckle imaging systems, where considerable benefits accrue from modest gains in the signal-to-noise ratio¹³. $\square$

Received 20 November 1990; accepted 29 March 1991.

1. Angel, J. R. P., Wizinowich, P., Lloyd-Hart, M. & Sandler, D. G. *Nature* **348**, 221–224 (1990).
2. Gardner, C. S. *Proc. IEEE* **77**, 408–418 (1989).
3. Hardy, J. W. *Proc. Soc. Photo-Instrum. Engrs* **1114**, 2–13 (1965).
4. Hulburd, B. & Sandler, D. G. *Opt. Engng* **29**, 1186–1190 (1990).
5. Hebb, D. O. *The Organization of Behavior* (Wiley, New York, 1949).
6. Hopfield, J. J. & Tank, D. W. *Proc. natn. Acad. Sci. U.S.A.* **79**, 2554–2558 (1982).
7. Kohonen, T. *Self-Organization and Associative Memory* (Springer, Berlin, 1984).
8. Rosenblatt, F. *Principles of Neurodynamics* (Spartan, Washington, DC, 1962).
9. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* Vol. 1, 318–362 (MIT Press, Massachusetts, 1986).
10. Born, M. & Wolf, E. *Principles of Optics* 6th Edn, 416 (Pergamon, Oxford, 1984).
11. Noll, R. J. *J. opt. Soc. Am.* **66**, 207–211 (1976).
12. Fried, D. L. *J. opt. Soc. Am.* **55**, 1427–1435 (1965).
13. Roddier, F., Northcott, M. & Graves, J. E. *Publ. astr. Soc. Pacif.* **103**, 131–149 (1991).

ACKNOWLEDGEMENTS. We thank J. R. P. Angel, P. Wizinowich, P. Hammerling and V. Smiley for many useful conversations, and J. Spinhrine (Rockwell Power Services) for setting up the CCD camera optics. This work was supported by the US Air Force Phillips Laboratory with contractual assistance from the Office of Naval Research.

Saccharide-sensitive phase transition of a lectin-loaded gel

Etsuo Kokufata*, Yong-Qing Zhang & Toyochi Tanaka

Department of Physics and Center for Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

A GEL system that swells and shrinks in response to specific molecules could serve as the basis for technological applications of polymer gels, for example as sensors, drug-delivery devices and actuators. Here we describe a lectin-loaded polymer gel that undergoes distinct swelling behaviour in response to different saccharides. The gel consists of a covalently cross-linked polymer network of *N*-isopropylacrylamide in which the lectin, concanavalin A, is immobilized. Concanavalin A displays selective binding affinities for certain saccharides. The gel undergoes a volume phase transition at $\sim 34^\circ\text{C}$. When the saccharide dextran sulphate is added (as the sodium salt DSS) to the gel, it swells to a volume up to five times greater at temperatures close to this transition, and the transition itself changes from discontinuous to continuous. Replacing DSS with the non-ionic saccharide α -methyl-D-mannopyranoside brings about collapse of the gel back to almost its native volume. This process is reversible and repeatable. These results point to a general principle for the design of such molecule-specific systems.

Many gels made of synthetic and natural polymers undergo reversible, discontinuous volume changes in response to

* Permanent address: Institute of Applied Biochemistry, University of Tsukuba, Tsukuba-shi, Ibaraki 305, Japan.

temperature¹⁻⁶, solvent or ionic composition^{2,5-7}, pH³, small electric fields⁹ and light^{10,11}. These variables are not specific: in the pH-driven phase transition, for example, any acids or bases will induce the transition³. Our aim was to develop a gel in which reversible swelling and shrinking is induced by a specific set of molecular stimuli.

Concanavalin A is a globular lectin¹² made of four subunits, each of dimensions $44 \times 40 \times 39$ Å. It has different binding affinities towards different saccharides¹³. Dextran sulphate sodium (DSS, relative molecular mass, 7,000, sulphur content 18 wt%) has a specific affinity for Con A, and, on binding to a Con A-containing gel, produces an excess ionic osmotic pressure in the gel that raises the volume phase transition temperature. Another saccharide, α -methyl-D-mannopyranoside (MP), also has a strong affinity for Con A¹² and is used as a competitive inhibitor for DSS-Con A association. Con A can be entrapped in a gel composed of *N*-isopropylacrylamide (NIPA), presumably by mechanical constraints. We selected this gel because of the availability of information^{4,6} on its volume phase transition. As temperature is varied within a certain range, the gel containing the Con A-DSS complex (DSS-gel) swells because of the ionic osmotic pressure exerted by DSS. When DSS is replaced by MP, the gel (MP-gel) shrinks from the swollen DSS form back to essentially the volume of the native gel, because the neutral MP exerts no ionic pressure (Fig. 1).

DSS-gel was prepared by first making the Con A-DSS complex and then immobilizing the complex in the NIPA gel. Using the Con A-DSS complex rather than free Con A provided a marked increase in the swelling of the gel, presumably because of protection against chemical alteration of the active site of the Con A protein during gelation. Con A (0.4 mg, Seikagaku Kogyo) and DSS (40 mg, Kowa) were dissolved in 1 ml of water. This was mixed with 1 ml of an aqueous solution containing NIPA (156 mg, polymer constituent, Kodak), *N,N'*-methylenebisacrylamide (2.66 mg, cross-linker, BioRad), and tetramethylethylenediamine (4.8 μ l, accelerator, BioRad). The solution was degassed, and 0.4 mg of ammonium persulphate (initiator, Mallinckrodt) was added to induce gelation. Gelation took place in glass capillaries with 0.1 mm inner diameter at 0°C. After 1 h, gelation was completed, and the gels were removed from the capillaries and washed thoroughly with water. The samples were cut into 2-mm-long cylinders and stored at 3°C before use.

We measured the volume changes of DSS- and MP-gels in water as a function of temperature. MP-gel was obtained by soaking the DSS-gel in an aqueous MP solution (0.1 M) for 1 h, followed by washing with water. Gel diameters were determined using a microscope with a calibrated scale, the temperature being controlled to within 0.1°C. MP-gel underwent a discontinuous phase transition at 34.2°C (Fig. 2). The swelling curve was similar to those of the pure NIPA gel without Con A and

that of the gel with free Con A. The transition temperatures of the pure NIPA gel⁶ and the gel with free Con A were the same, 33.8°C. In contrast, when DSS ions were bound to Con A, the gel underwent a sharp but continuous volume change at 34.8°C, 0.6°C higher than the transition temperature of MP-gel. This difference is significant and reproducible.

Control of swelling and shrinking of the lectin-loaded gel at 34.5°C was checked by the repeated, alternate binding of DSS and MP to the gel-immobilized Con A. MP-gel was converted into DSS-gel by soaking the gel in 2 wt% DSS solution for 1 h, then again converted into MP-gel in the same manner as described above. The gel swelled to five times its original volume when DSS was bound to Con A, whereas the replacement of DSS by MP brought about collapse of the gel. We could repeat these swelling and shrinking cycles with excellent reproducibility.

The swelling and shrinking changes of Con A-loaded NIPA gel are presumably determined by a balance between the pressure due to repulsive electrostatic interactions of ionized dextran sulphate (DS^{n-}) molecules, the osmotic pressure owing to counter ions (Na^+) from DSS, and the attractive hydrophobic interaction among NIPA residues. The first two factors act to raise the transition temperature. It is interesting that the phase transition, discontinuous for the MP-gel and the gel with free Con A, becomes continuous when DSS is bound to Con A. The introduction of ionizable groups to the gel network by copolymerization has been known to make the transition discontinuous and shift it towards larger volume changes⁴. In the present system, the DSS molecules have ionized sulphate groups which bind to the lectin entrapped in the network. This may cause a localized raise in ion concentration within the gel phase, and thus a range of 'local' transition temperatures, resulting in a continuous transition.

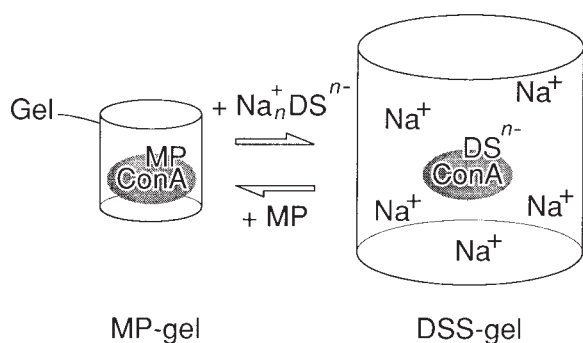


FIG. 1 Schematic illustration of saccharide-responsive, reversible swelling of a *N*-isopropylacrylamide gel loaded with concanavalin A. $Na^+ DS^{n-}$ is dextran sulphate sodium (DSS).

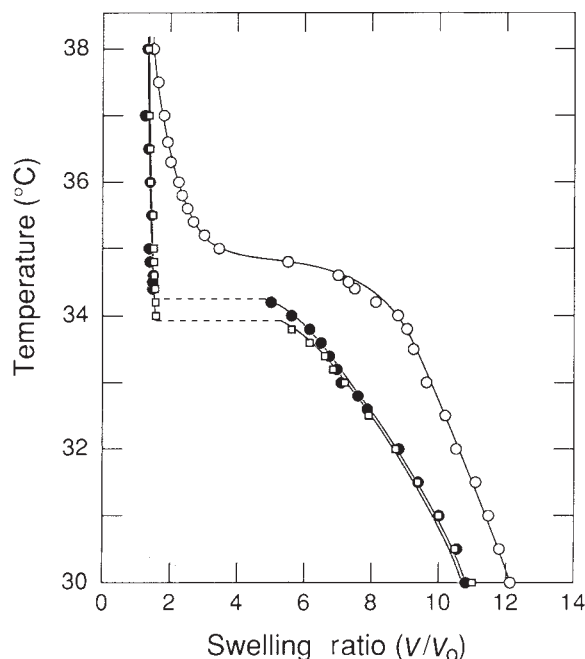


FIG. 2 Temperature dependence for equilibrated volumes of NIPA gel including the Con A-DSS complex (DSS-gel, ○), MP (MP-gel, ●), and free of both DSS and MP (□). The latter was prepared as a control sample as described in the text but using an aqueous Con A solution instead of the Con A-DSS solution. Hysteresis was observed in the volume changes of DSS-gel and the free-Con A gel on heating and cooling, indicating a discontinuous phase transition. The diameter of each gel in the collapsed state, determined at 50°C, was $d_0 = 0.074$ mm; the volume of this gel is denoted V_0 . The concentration of dry matter in the collapsed state was estimated from the preparation recipe to be 90 wt%.

Certain other saccharides, to which ionizable groups such as $-NH_2$, $-COOH$ and $-OP(=O)(OH)_2$ are chemically bound, are currently available. Various lectins have been isolated and characterized in terms of their interaction with saccharides. Therefore, different combinations of saccharide-lectin systems are available for designing saccharide-sensitive gels. More generally, to design a gel that undergoes a discontinuous volume transition in response to a molecular stimulus may be achieved by embedding an active element into the gel that interferes with gel equilibrium in response to the presence of the 'stimulus' molecule. □

Received 2 January; accepted 18 March 1991.

1. Tanaka, T. *Phys. Rev. Lett.* **40**, 820-823 (1978).
2. Dusek, K. & Patterson, D. *J. Polym. Sci.* **A26**, 1209-1216 (1968).
3. Tanaka, T. *et al. Phys. Rev. Lett.* **45**, 1636-1639 (1980).
4. Ilavsky, M. *Macromolecules* **15**, 782-788 (1982).
5. Hirokawa, Y. & Tanaka, T. *J. chem. Phys.* **81**, 6379-6380 (1984).
6. Hirokawa, Y., Tanaka, T. & Katayama, S. in *Microbial Adhesion and Aggregation* (ed. Marshall, K.C.) 177 (Springer, Berlin, 1984).
7. Hrouz, J., Ilavsky, M., Ulbrich, K. & Kopecek, J. *Europ. Polym. J.* **17**, 361-366 (1982).
8. Ohmine, I. & Tanaka, T. *J. chem. Phys.* **77**, 5725-5729 (1982).
9. Tanaka, T., Nishio, I., Sun, S.-T. & Ueno-Nishio, S. *Science* **218**, 467-469 (1982).
10. Mamada, A., Tanaka, T., Kungwachakun, D. & Irie, M. *Macromolecules* **23**, 1517-1519 (1990).
11. Suzuki, A. & Tanaka, T. *Nature* **346**, 345-347 (1990).
12. Wang, J. L., Cunningham, B. A. & Edelman, G. M. *J. biol. Chem.* **250**, 1513-1521 (1975).
13. Goldstein, I. J., Hollerman, C. E. & Smith, E. E. *Biochemistry* **4**, 876-883 (1965).

ACKNOWLEDGEMENTS. This work was supported by the NSF. The stay of E.K. in the United States was supported by the Japanese Ministry of Education, Science and Culture. DSS was supplied by Kowa Company through the kind assistance of H. Tani.

Model estimates of CO₂ emissions from soil in response to global warming

D. S. Jenkinson*, D. E. Adams & A. Wild

Department of Soil Science, University of Reading, Reading RG1 5AQ, UK
* Also at Rothamsted Experimental Station, Harpenden AL5 2JQ, UK

ONE effect of global warming will be to accelerate the decomposition of soil organic matter, thereby releasing CO₂ to the atmosphere, which will further enhance the warming trend¹⁻⁷. Such a feedback mechanism could be quantitatively important, because CO₂ is thought to be responsible for ~55% of the increase in radiative forcing arising from anthropogenic emissions of gases to the atmosphere⁸, and there is about twice as much carbon in the top metre of soil as in the atmosphere⁹. Here we use the Rothamsted model for the turnover of organic matter in soil³ to calculate the amount of CO₂ that would be released from the world stock of soil organic matter if temperatures increase as predicted, the annual return of plant debris to the soil being held constant. If world temperatures rise by 0.03 °C yr⁻¹ (the increase considered as most likely by the Intergovernmental Panel on Climate Change⁸), we estimate that the additional release of CO₂ from soil organic matter over the next 60 years will be 61 × 10¹⁵ gC. This is ~19% of the CO₂ that will be released by combustion of fossil fuel during the next 60 years if present use of fuel continues unabated.

Climate change can influence the stock of soil organic matter in two ways: by altering plant growth, thus altering the annual return of plant debris to the soil; and by changing the rate at which this input decays in or on the soil. Here we are concerned with the second of these processes: the effects of climate change, particularly global warming, on the decay process.

The current Rothamsted model³ was developed to model the effects of agricultural management on the stock of organic matter in soil, which it does reasonably accurately over timescales of years to a century¹⁰. Here we extend the model to the global

TABLE 1 Stock of carbon and calculated annual input of carbon for the world's soils

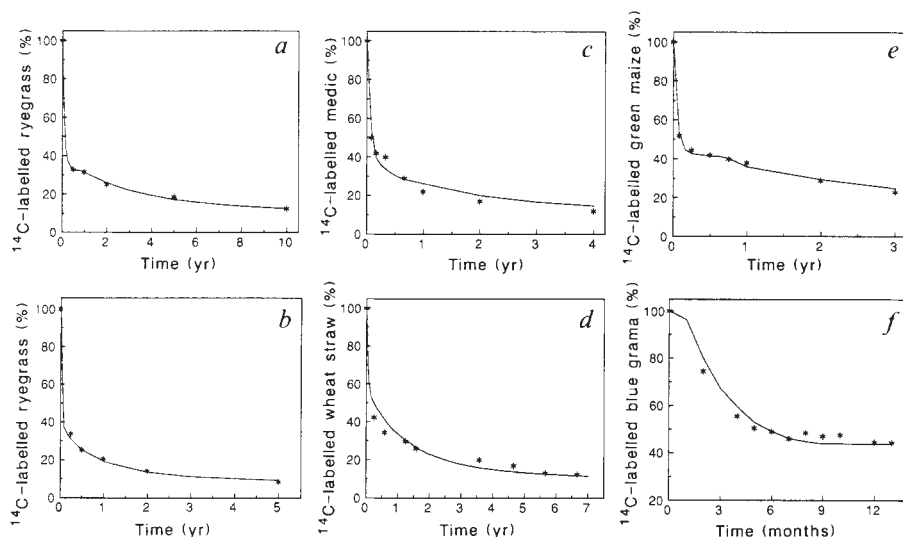
Life zone	Carbon stock (×10 ¹⁵ g)	Carbon input (×10 ¹⁵ g yr ⁻¹)
Tundra	191	0.9
Boreal desert	20	0.1
Cool desert	43	0.9
Warm desert	20	0.6
Tropical desert bush	2	0.1
Cool temperate steppe	120	2.7
Temperate thorn steppe	30	1.8
Tropical woodland and savanna	129	11.5
Boreal forest moist	49	0.8
Boreal forest wet	133	4.7
Temperate forest cool	43	3.1
Temperate forest warm	61	7.1
Tropical forest very dry	22	1.7
Tropical forest dry	24	1.1
Tropical forest moist	60	13.2
Tropical forest wet	78	15.3
Cultivated land	167	10.2
Wetlands	202	—
Total	1,394	75.8

Carbon stock is from ref. 9. Carbon input is calculated from Rothamsted carbon model, assuming all soils contain 20% clay and 300 g C m⁻² as inert organic matter, and have reached carbon equilibrium under the vegetation characteristic of each life zone. The DPM/RPM ratio is taken to be 0.25 for all the forests, 0.43 for tropical woodland and savanna, 0.5 for the deserts, tundra and cool temperate steppe, and 0.67 for cultivated land and temperate thorn steppe. Weather data (mean monthly air temperature in °C, monthly rainfall in mm and monthly potential evapotranspiration in mm, all averaged over at least 10 years), were taken from ref. 25, as means of 4-16 sites in each of the 16 life zones, allowing for phase differences between Northern and Southern Hemispheres. Cultivated land is assumed to have the climate of the life zone in which it occurs.

scale and use it to predict how much CO₂ will be liberated from the world stock of soil organic matter for a given rise in temperature. This five-compartment model has two input compartments: decomposable plant material (DPM) and resistant plant material (RPM), both of which decompose, by first-order processes which have characteristic (and different) rates, to CO₂ (lost from the system), microbial biomass and humified organic matter. Both microbial biomass and humified organic matter decompose at their characteristic rates by first-order processes to give more CO₂, biomass and humified matter. The soil is also assumed to contain a small organic compartment that is inert to biological attack. The model allows for the effects of temperature, soil moisture content, plant cover (decomposition being faster in bare soil than under vegetation¹¹) and soil clay content on the decomposition processes. It works in monthly intervals, using mean monthly air temperatures. Soil moisture is updated monthly from the rainfall and evaporation of the preceding month. The model parameters were set by tuning to data from the Rothamsted long-term field experiments, for timescales of decades to centuries¹⁰, and to experiments on the decay of ¹⁴C-labelled ryegrass for timescales of months to decades.^{11,12} With the exception of the DPM/RPM ratio (see legend to Fig. 1), none of the parameters thus set were altered in the calculations presented here.

We first tested the model to see how well it could predict the decomposition of plant material in soil in a range of climates and soils, using data from experiments¹¹⁻¹⁸ on the rate of decomposition of plant material uniformly labelled with ¹⁴C in the field. Data from 31 experiments, 19 in the humid tropics, 6 in the temperate zone, 4 in the warm temperate forest zone and 2 in the cold temperate steppe zone, were available: Fig. 1 shows the fit between model and measurement for 6 of them. The results in Fig. 1a and b do not provide an independent check of the validity of this model, as these particular results had

FIG. 1 Fits between modelled and measured data (solid line and stars, respectively) for the decomposition of plant material uniformly labelled with ^{14}C at 6 sites under field conditions. *a*, Rothamsted, UK; *b*, Ibadan, Nigeria; *c*, Northfield, Australia; *d*, Patarra, Costa Rica; *e*, Vienna, Austria; *f*, Pawnee, USA. The vertical axis shows the percentage of carbon remaining. In fitting the data, all the model parameters were as set previously³, except for the ratio of decomposable plant material (DPM) to resistant plant material (RPM) in the incoming plant debris. The DPM/RPM ratio was first optimized to obtain the best fit in each of the 27 data sets: the DPM/RPM ratios thus found were aggregated to give a mean for each type of plant material. For example, the mean DPM/RPM ratio used for wheat straw in (*d*) was 1.44, a mean for 13 experiments. The labelled plant materials were ryegrass leaves and stems^{11,12} (*a* and *b*), medic leaves¹⁵ (*c*), wheat straw¹⁶ (*d*), green maize leaves¹³ (*e*) and blue grama grass tops¹⁴ (*f*). Weather data are long-term means from the nearest meteorological site.



already been used in setting the model parameters. Excluding these and one other calibration site, the model accounted for 83% of the variance in the measured data from the 28 remaining sites. The worst fits were for two allophanic soils¹⁶, in which decomposition was slower than modelled, and for the two Canadian sites¹⁸, in which decomposition was much faster than modelled (and indeed as measured¹⁶ in tropical sites using the same plant material, wheat straw). Excluding these, the model accounted for 90% of the variance in the remaining 24 experiments. Experiments on the decomposition of labelled plant material under field conditions are particularly valuable in model testing, as they give reliable measures of carbon retention over much longer decomposition periods than, for example, litter-bag experiments. Unfortunately, relatively few have been done, and so far only a narrow range of labelled plant materials, mostly annuals, has been used.

We then used the model to calculate how much carbon would have to be returned to the land surface of the Earth each year to maintain the present stock of soil organic carbon in the present climate. For this, we divided the parts of the globe that have vegetation into 16 climate zones (plus an additional 'cultivated land'), based on the aggregation by Post *et al.* of Holdridge's¹⁹ 'life zones'. Table 1 shows the carbon contents of the different zones, taken from Post *et al.*⁹ and the calculated annual input for each zone. Each zone was assumed to have attained equilibrium under its prevailing (and constant) climate and vegetation; the model then calculated the (constant) annual input

required to maintain the carbon content of that zone. To do this, the model was run for 10,000 years (20,000 for the tundra), under the climate conditions pertaining to a particular life zone, varying the annual input of plant material iteratively until this input generated the steady-state soil organic carbon content of that zone as given in Table 1. The specification that climates and inputs are unchanged for these long periods is less restrictive than might appear: the model parameters are such that climate or input changes occurring more than 250 years ago have little effect on the calculated input for all except the tundra zone.

The aggregated return of carbon for all zones as calculated by the model is $75.8 \times 10^{15} \text{ g C yr}^{-1}$, a little greater than published⁸ estimates of net primary production (NPP), which are $\sim 60 \times 10^{15} \text{ g C yr}^{-1}$. Wetlands are excluded from this calculation because the Rothamsted model is not applicable to anaerobic soils. NPP and calculated carbon inputs to soil should be the same if all the vegetation ultimately decomposes in or on the soil. Losses by burning (estimated²⁰ to be $2\text{--}5 \times 10^{15} \text{ g C yr}^{-1}$) should be deducted from NPP, making the discrepancy greater. But even this rough agreement between NPP and calculated input is encouraging, considering the very different ways the two figures are obtained. The discrepancy could be eliminated by reducing the rate constants in the model, but these should be (and are) set independently. The main source of error in calculating carbon input from soil data is probably the presence of biologically inert carbon. We set this everywhere at 300 g C m^{-2} , but it could often be much greater, particularly for

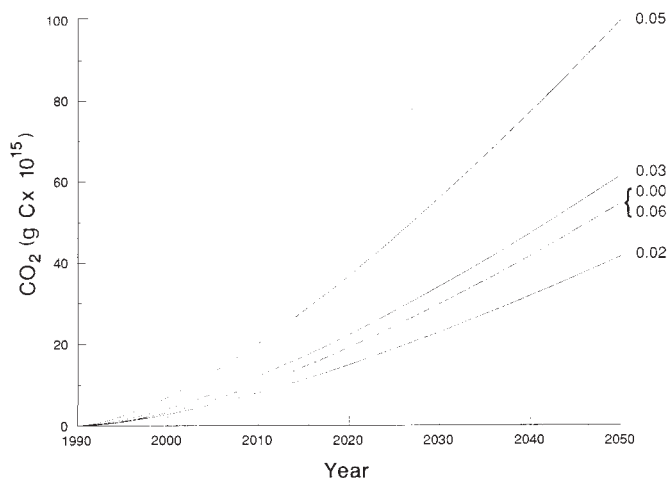


FIG. 2 Carbon released as CO_2 from soil organic matter. Numbers on right-hand side give annual temperature increase ($^{\circ}\text{C}$). Solid lines: net carbon loss from the world stock of soil organic matter as a result of a uniform 0.02, 0.03 and $0.05 \text{ }^{\circ}\text{C yr}^{-1}$ rise in global temperature in each of the 17 life zones in Table 1, excluding wetlands. The rise in temperature was assumed to be the same throughout the year. Broken line: net carbon loss as a result of a $0.03 \text{ }^{\circ}\text{C yr}^{-1}$ average rise in global temperature, divided unequally between the different life zones, with a rise of $0.06 \text{ }^{\circ}\text{C yr}^{-1}$ in the high-latitude zones (tundra, boreal desert, cool desert, cool temperate steppe, moist boreal forest, wet boreal forest and cool temperate forest, which together contain 50% of the total non-wetland soil organic carbon) and no rise in the 10 remaining life zones; again excluding wetlands.

heavy clay soils and for the deeper soil horizons (the soil depth used by Post *et al.* was 1 m). The setting of clay content at 20% for all soils is another obvious source of error: if, for example, the clay content of the tropical forest wet zone is set at 10, 20 and 30%, the calculated inputs of carbon for equilibrium in this zone are 17.3, 15.3 and 14.4×10^{15} g C yr⁻¹, respectively.

The soil carbon estimates in Table 1 do not include surface litter: if this is included, the discrepancy between calculated input and global NPP will increase, although the effect is small. Consider the worst case, that of boreal forest. If, following Schlesinger²¹, boreal forest contains 13% of its total soil organic carbon in surface litter, then the calculated annual input of carbon from the (combined) boreal forest moist and wet zones, allowing for surface litter, is 6.3×10^{15} g C yr⁻¹, not 5.5×10^{15} as in Table 1.

In the model the decomposability of the incoming plant material is set by the DPM/RPM ratio and errors will result if this ratio is incorrect. The value of 0.25 used in Table 1 for all the forests is taken from work on deciduous trees at Rothamsted (D.S.J., D. D. Harkness, E. D. Vance, D.E.A. and A. F. Harrison, manuscript in preparation) and may be too large for the more resistant debris from coniferous forest. If so, the calculated annual inputs of carbon given in Table 1 for the boreal forests are too big.

As the model gave plausible results on the world scale, we then used it to calculate how much CO₂ would be released from the world stock of soil organic carbon if world temperatures rose uniformly by 0.2, 0.3 or 0.5 °C per decade. These are the lower limit, best estimate and upper limit suggested by the IPCC report⁸ for global warming over the next decades. The annual inputs of carbon to the life zones are assumed to be as in Table 1 and to continue unchanged. Figure 2 shows that the extra CO₂ carbon evolved in 60 years from soil organic carbon as a result of a rise of 0.02 °C yr⁻¹ is 41×10^{15} g; for 0.03 °C yr⁻¹, 61×10^{15} g and for 0.05 °C yr⁻¹, 100×10^{15} g. The corresponding evolution of CO₂ carbon from combustion of fossil fuel over the same 60-yr period will be 320×10^{15} g, if present production⁸ (5.4×10^{15} g yr⁻¹) continues unchanged. If warming is confined to northern latitudes, less soil organic carbon will be decomposed than if warming is uniform (Fig. 2). This is because at high latitudes the soil is too cold for appreciable biological activity for much of the year: if the mean monthly temperature for January is -15 °C, an increase of 10 °C will have no effect on decomposition. These calculations suggest that increased decomposition of soil organic carbon will indeed make an important contribution to the greenhouse effect, but that the contribution is unlikely to be a runaway effect, as has been feared².

The projections in Fig. 2 are very uncertain: individual predictions could be in error by $\pm 50\%$. The fits illustrated in Fig. 1 show that the parts of the model which adjust decomposition rates for climate and soil texture are reasonably satisfactory: the model accounts for 83% of the variance in the unselected database. But this base is itself biased by the uneven geographical spread of the experiments, and some life zones are completely unrepresented. Even more serious errors are introduced when the model is fitted to the global data in Table 1: the global stock of soil organic matter itself has a coefficient of variation⁹ of $\sim 14\%$.

The model can also be used to see how changes in precipitation influence decomposition, although on the global scale this must remain an academic exercise until better forecasts of the effect of increasing concentrations of greenhouse gases on precipitation are available. According to the model, in the cool temperate steppe zone, a 25% decrease in precipitation, spread uniformly over the year and starting in 1990, will increase the stock of soil organic carbon from 120×10^{15} g C (see Table 1) to 125×10^{15} g by 2050, the annual input being unchanged throughout. In this zone, lack of moisture retards decomposition during summer, when it would otherwise have been at its fastest. In contrast,

neither a 25% increase nor a 25% decrease in rainfall alters the stock of soil organic carbon in the tropical forest wet zone; the soil is always wet enough for decomposition to proceed at its maximum permitted rate.

This work shows how climate change can influence the decomposition process in soil. But the amount of organic carbon present in soil at a particular time is determined both by the rate of decomposition and by the annual input of plant debris, set constant in Table 1 and Fig. 2. To tackle the much larger question of how climate change affects the flux of CO₂ between soil and atmosphere²²⁻²⁴, our model would have to be coupled to another that specifies how global plant production (and hence annual return of plant debris to soil) is altered by increasing concentrations of greenhouse gases. □

Received 5 December 1990; accepted 11 April 1991.

- Schleser, G. H. Z. *Naturf.* **A37**, 287-219 (1981).
- Kohlmair, G. H., Janacek, A. & Kindermann, J. in *Soils and the Greenhouse Effect* (ed. Bouwman, A. F.) 415-422 (Wiley, Chichester, 1990).
- Jenkinson, D. S. *Phil. Trans. R. Soc. Lond.* **B329**, 361-368 (1990).
- Lashof, D. A. *Clim. Change* **14**, 213-242 (1989).
- Schimel, D. S., Parton, W. J., Kittel, T. G. F., Ojima, D. S. & Cole, C. V. *Clim. Change* **17**, 13-25 (1990).
- Tinker, P. B. & Ineson, P. in *Soils on a Warmer Earth* (eds Scharpenseel, H. W., Schomaker, M. & Ayoub, A.) 71-87 (Elsevier, Amsterdam, 1990).
- Franz, E. H. in *Soils on a Warmer Earth* (eds Scharpenseel, H. W., Schomaker, M. & Ayoub, A.) 109-120 (Elsevier, Amsterdam, 1990).
- Intergovernmental Panel on Climate Change (eds Houghton, J. T., Jenkins, G. J. & Ephraums, J. J.) (Cambridge University Press, 1990).
- Post, W. M., Emanuel, W. R., Zinke, P. J. & Stangenberger, A. G. *Nature* **289**, 156-159 (1982).
- Jenkinson, D. S., Hart, P. B. S., Rayner, J. H. & Parry, L. C. *INTECOL Bull.* **15**, 1-8 (1987).
- Jenkinson, D. S. *J. Soil Sci.* **28**, 424-434 (1977).
- Ayanaba, A. & Jenkinson, D. S. *Soil Sci. Soc. Am. J.* **54**, 112-115 (1990).
- Oberlander, H. E. *Scr. varia pont. Acad. Sci.* **38**, 1001-1062 (1973).
- Nyhan, J. W. *Soil Sci. Soc. Am. J.* **39**, 643-648 (1975).
- Ladd, J. N., Oades, J. M. & Amato, M. *Soil Biol. Biochem.* **13**, 119-126 (1981).
- Gonzalez, A. M. A. & Sauerbeck, D. R. in *Regional Colloquium on Soil Organic Matter Studies* (eds Cerri, C. C., Athie, D. & Sodrzejewski, D.) 141-146 (CENA/USP & PROMOCET, Piracicaba, 1982).
- Neue, H. U. & Scharpenseel, H. W. *Soil Sci. Total Envir.* **62**, 431-434 (1987).
- Voroney, R. P., Paul, E. A. & Anderson, D. W. *Can. J. Soil Sci.* **69**, 63-77 (1989).
- Holdridge, L. R. *Life Zone Ecology* (Tropical Science Center, San Jose, 1964).
- Crutzen, P. J. & Andreae, M. O. *Science* **250**, 1669-1678 (1990).
- Schlesinger, W. H. *Ann. Rev. ecol. Syst.* **8**, 51-81 (1977).
- Adams, J. W., Faure, H., Faure-Denard, L., McGlade, J. M. & Woodward, F. I. *Nature* **348**, 711-714 (1990).
- Prentice, K. C. & Fung, I. Y. *Nature* **346**, 48-51 (1990).
- Schlesinger, W. H. *Nature* **348**, 232-234 (1990).
- Müller, M. J. *Selected Climatic Data for a Global Set of Standard Stations for Vegetation Science* (Junk, The Hague, 1982).

ACKNOWLEDGEMENTS. We thank D. Barraclough for advice on computer programming, J. I. L. Morison for help in locating climatic data and the Oxford University Computing Service for access to computing facilities. This work was supported by the Leverhulme Trust.

Water diffusion in a basaltic melt

Youxue Zhang & E. M. Stolper

Division of Geological and Planetary Sciences, 170-25,
California Institute of Technology, Pasadena, California 91125, USA

WATER is the most abundant volatile component in terrestrial basalts and is a significant constituent of the gases that escape from basaltic magmas. Knowledge of the diffusivity of water (and other volatiles) in basaltic melts is important for understanding the degassing of basaltic magma and for assessing the fractionation of volatiles during degassing. We report here measurements of water diffusivity in a basaltic liquid. The water concentration profiles through the samples, determined by Fourier-transform infrared spectroscopy, cannot be modelled adequately on the basis of a constant water diffusivity¹⁻⁷, but instead can be fitted by assuming that only molecular H₂O is diffusing and that there is a local equilibrium between H₂O molecules and OH groups⁷⁻¹³. The concentration-dependent total water diffusivities in the basaltic melt at 1,300-1,500 °C are 30-50 times as large as those in rhyolitic melts⁴⁻⁷, and are greater than the total CO₂ diffusivity in

basaltic melts, contrary to previous expectations¹⁴. These results suggest that diffusive fractionation would increase the ratio of water to carbon dioxide in growing bubbles relative to equilibrium partitioning, and decrease the ratio in interface melts near an advancing anhydrous phenocryst.

Chips of basaltic glass from the Juan de Fuca (JDF) ridge were used as starting material for this study. The original glass chips contained 0.35–0.43 wt% dissolved total water (hereafter % refers to wt% unless otherwise specified) as measured by FTIR (Fourier transform infrared spectroscopy; see also ref. 15), and some small bubbles (<100 μm) and microphenocrysts (<200 μm), one hundred to several hundred micrometres apart. Water-poor glass was prepared by melting ~10 grams of the JDF basaltic glass powder (contained in a platinum crucible pre-saturated with FeO-bearing basaltic melt) in a 1-atm Deltech furnace at a f_{O_2} corresponding to the quartz-fayalite-magnetite buffer. Glass prepared in this way contains 0.02–0.10% total water (based on FTIR measurement) and is almost free of bubbles and crystals. The partially dehydrated glass is more

transparent than the original, more water-rich glass. Even though the water contents of the dehydrated and the original starting glass material are variable, chips the size of those used in experiments (~2 mm in each dimension) are assumed to be uniform (within 5% relative) in water content. This assumption is based on analyses of selected fragments of the glasses.

Each experimental couple consisted of two halves, one a piece of pristine JDF glass and the other a piece of dehydrated JDF glass. Each half was a polished cylindrical disk ~1.5-mm thick and ~2 mm in diameter. The two halves fitted snugly into a graphite capsule, which was enclosed in a platinum capsule and then welded shut. The platinum capsule was placed in a sleeve of crushable MgO and then encased in a 30-mm-long graphite heater. CaF_2 was used as the pressurizing medium. Experiments were conducted in a piston-cylinder apparatus at 1,300–1,500 °C and 1.0 GPa. (An additional experiment was run at 1,250 °C, but the charge partially crystallized.) During the run, the interface between the two halves was roughly horizontal and the water-rich half was above the water-poor half. We used a 'piston-

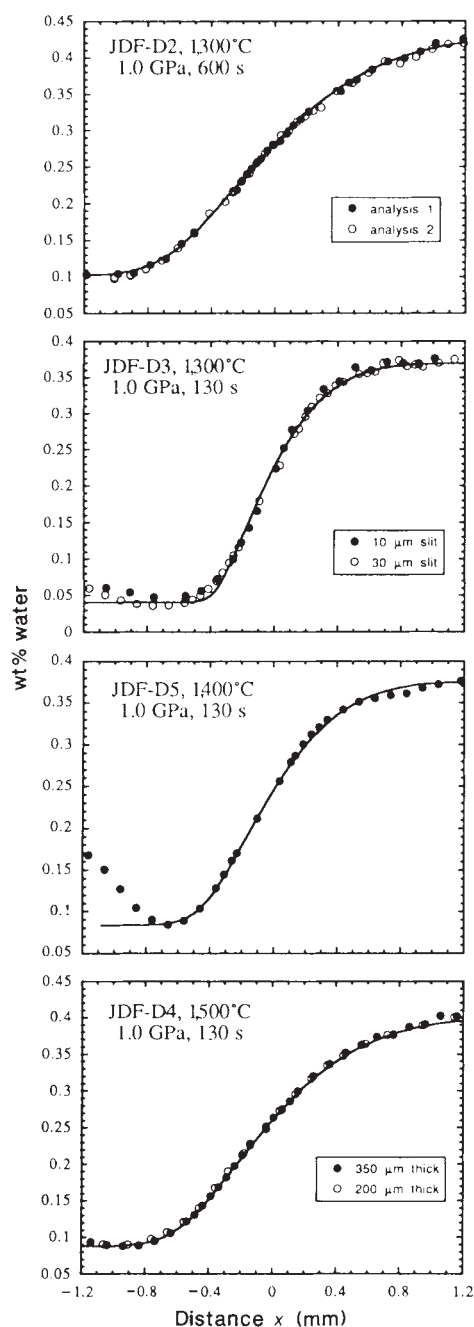


FIG. 1 Total water concentration profiles (points) with run conditions. The curves shown are fits of the data with a water diffusion model in which only molecular H_2O diffuses and there is local equilibrium between H_2O molecules and OH groups. K for all fits was arbitrarily chosen as 0.5. The $D_{\text{H}_2\text{O}}$ values derived for the four diffusion couples were $(3.23 \pm 0.5) \times 10^{-9}$, $(5.1 \pm 0.2) \times 10^{-9}$, $(7.3 \pm 0.2) \times 10^{-9}$ and $(1.20 \pm 0.02) \times 10^{-8} \text{ m}^2 \text{ s}^{-1}$ for JDF-D2, D3, D5 and D4 respectively. The values probably do not correspond to the actual diffusivity of molecular H_2O , as K was estimated by extrapolation from rhyolitic glasses. When K was varied, the quality of the fit changed little, but the best-fit $D_{\text{H}_2\text{O}}$ values varied roughly proportionately with K (refs 7, 11, 13). $D_{\text{H}_2\text{O}}$ values for JDF-D2 and JDF-D3 agree within 44%. All fits were done on a molar basis and then recalculated to wt% for plotting. Bulk composition of the basalt (TT152-21) by wt% is: SiO_2 , 50.6; TiO_2 , 1.88; Al_2O_3 , 13.9; $\text{FeO}_{\text{total}}$, 12.5; MnO , 0.23; MgO , 6.56; CaO , 11.4; Na_2O , 2.64; K_2O , 0.17; P_2O_5 , 0.21 (S. van der Laan, personal communication, and ref. 15). Water concentrations were determined from infrared spectra obtained with a Nicolet 60SX FTIR. A slit ($10 \mu\text{m} \times 2 \text{ mm}$ or $30 \mu\text{m} \times 1 \text{ mm}$) was used to delimit the infrared beam. The profiling was done with a translation stage⁷. Total water concentration was determined from the intensity of the absorption band at $\sim 353 \text{ cm}^{-1}$ using a molar absorptivity of $63 \text{ l mol}^{-1} \text{ cm}^{-1}$ (ref. 15). Concentrations of individual species, H_2O molecules and OH groups, could not be determined because the water contents were low. Baselines were drawn with a french curve. A small correction (less than 5% relative, depending on the total water concentration) to the peak at $\sim 353 \text{ cm}^{-1}$ was sometimes applied for the growth of an ice band in the detector at $\sim 325 \text{ cm}^{-1}$ during the measurements. From the quality of the spectra and reproducibility shown in this figure, the 1σ analytical precision in determining total water concentration is $\leq 2\%$ at $\geq 0.3\%$ water and $\sim 4\%$ at 0.05% water, less than or equal to the size of the points. The accuracy of the water concentration measurement (which depends on the chosen value of molar absorptivity) does not affect the determination of total water diffusivities. Nor does it affect the model fitting shown here, although it affects actual $D_{\text{H}_2\text{O}}$ values. The interface shown here is the Matano interface²⁰ determined from the smoothed concentration profile (actual interface positions were not measured). There are two complications to the concentration profiles. One is the presence of cracks, which can interfere with infrared analyses and complicate distance determinations. Sample JDF-D3 was affected the most by these cracks. The other is a (usually slight) increase in water content towards the end of the water-poor half, possibly due to water in pore spaces in the graphite capsule. The capsule of JDF-D5 was accidentally dropped into water before the platinum capsule was completely sealed. It was subsequently heated in a vacuum furnace at $\sim 110^\circ \text{C}$ for about half an hour and then welded shut, but is probable that water got into the capsule and was not totally removed. The increase towards the end of the water-poor half was ignored during fitting.

out' procedure (initially overpressurized to ~1.5 GPa) to bring the charge to 1.0 GPa cold, and then heated it. Temperatures were monitored and controlled by a Pt-Pt₉₀Rh₁₀ thermocouple connected to an automatic control system. Overshooting and fluctuation in temperature were $\leq 5^\circ\text{C}$ and $\sim 1^\circ\text{C}$ respectively. The temperature difference from the centre to the end of the charge is $\sim 10^\circ\text{C}$ (ref. 16 and M. B. Baker, personal communication). Experimental durations were 2–10 min. These short run durations were dictated by the high water diffusivities at these temperatures and the short diffusion distance available, and resulted in large relative errors in determining the experiment duration, as the heating took ~ 0.5 min. To assess this error, a 'zero-time' experiment was carried out at $1,300^\circ\text{C}$ (the sample was brought up to temperature, then quenched instantly); the concentration profile for this experiment is equivalent to diffusion at $1,300^\circ\text{C}$ for 8 s. Reported run durations (Fig. 1) are nominal run duration plus 10 s to account for the finite times of heating and quenching.

Each charge was quenched by shutting off the power to the piston-cylinder apparatus. A doubly polished, 200–350- μm -thick slice parallel to the axis of the cylinder was prepared near the widest section of the cylinder. Optical examination showed the charges to be free of crystals and bubbles (except one bubble $\sim 100\ \mu\text{m}$ in diameter in the water-rich side of JDF-D4, probably a residue of a large bubble in the initial glass), suggesting that microphenocrysts and bubbles in the original water-rich glass wafers dissolved into the melt during the experiments. Dissolution of bubbles and microphenocrysts had little effect on the water concentration profile, but it had a significant effect on the carbonate concentration profile (discussed later) because the dissolution doubled the CO_2 content of the glass. The degree of transparency varied across the charge, the water-rich side being less transparent as in the original glass chips.

Figure 1 shows the experimental conditions, analytical procedures and resulting total water concentration profiles. The diffusion medium is assumed to be infinite. If the profiles could be characterized by a constant total water diffusivity, they would be symmetric about the interface and fitted well by error functions, but Fig. 1 shows that the concentration profiles are not symmetric. When the profiles are fitted to an error function, there are systematic discrepancies between the data and the fit, except for JDF-D3. The inability of a constant $\bar{D}^*_{\text{water}}$ (average total water diffusivity⁷) to account in detail for water diffusion profiles in silicate melts and glasses is well known^{1–7}. The $\bar{D}^*_{\text{water}}$ values thus obtained (2.5×10^{-10} , 2.9×10^{-10} , 4.7×10^{-10} and $8.4 \times 10^{-10}\ \text{m}^2\ \text{s}^{-1}$ for JDF-D2, JDF-D3, JDF-D5 and JDF-D4), although imperfectly fitting the concentration profiles, characterize the mean transport rate of water under these conditions. The two experiments at $1,300^\circ\text{C}$ (JDF-D2 and JDF-D3), had run durations that differed by a factor of 5, but their $\bar{D}^*_{\text{water}}$ values agree to within 20% (formal 1σ errors for $\bar{D}^*_{\text{water}}$ values of both samples based on fits of profiles are $\sim 5\%$), demonstrating diffusion control of the experiments and providing an estimate of uncertainties in $\bar{D}^*_{\text{water}}$.

We also tested the data against a model in which only H_2O molecules diffuse^{7–13}, although the dominant hydrous species at these low water contents is expected to be OH. Local equilibrium was assumed¹⁷, with an equilibrium constant $K = 0.5$ chosen for the reaction



where O_{dry} refers to anhydrous oxygen¹⁸. The water concentration profiles were fitted following the procedure of Zhang *et al.*⁷ and the resulting fits are given in Fig. 1. In general, good fits are obtained, suggesting that molecular H_2O is the diffusing species for water diffusion in the basaltic melt, as found previously for silica^{8,12} and rhyolitic glasses⁷.

The variation of $\bar{D}^*_{\text{water}}$ (apparent total water diffusivity⁷) with total water content was calculated for each experiment from the best fit values of $D_{\text{H}_2\text{O}}$ (diffusivity of molecular H_2O) for a given

value of K by using the following expression^{7,19}

$$\bar{D}^*_{\text{water}} = D_{\text{H}_2\text{O}} d[\text{H}_2\text{O}]/d[\text{water}], \quad (2)$$

where quantities in brackets represent mole fractions and $[\text{water}] = [\text{H}_2\text{O}] + 0.5[\text{OH}]$ (ref. 18). Calculated values of $\bar{D}^*_{\text{water}}$ are shown as a function of water content in Fig. 2a. The values of $\bar{D}^*_{\text{water}}$ are almost proportional to total water content at these low water contents^{7,8,11}. We have also calculated $\bar{D}^*_{\text{water}}$ values using a Boltzmann-Matano analysis²⁰ of the concentration profiles, and these $\bar{D}^*_{\text{water}}$ values are compared with those calculated from (2) in Fig. 2a. They are in reasonable agreement, but $\bar{D}^*_{\text{water}}$ values obtained using the Boltzmann-Matano analysis show more scatter because errors arise from uncertainties in evaluating numerical derivatives and integrals.

The partially dehydrated starting glass chips were also partially decarbonated, and there is therefore a carbonate profile between the two halves of the diffusion couple. The graphite capsule also provides a source of carbonate to the basaltic melt (from oxidation of graphite by ferric iron) which diffuses inward into the melt. These carbonate concentration profiles were determined with the FTIR for JDF-D2. We only consider the profile adjacent to the graphite capsule in the initially decarbonated half because other profiles were affected considerably by dissol-

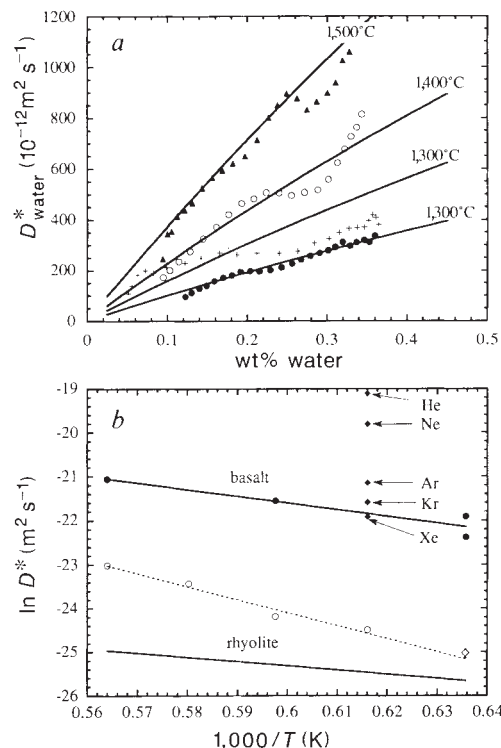


FIG. 2 a, Comparison of $\bar{D}^*_{\text{water}}$ values obtained from Boltzmann-Matano analysis (points) with those calculated from equation (2) based on the fits in Fig. 1 (solid lines). $\bullet$, $1,300^\circ\text{C}$ (JDF-D2); $+$, $1,300^\circ\text{C}$ (JDF-D3); $\circ$, $1,400^\circ\text{C}$ (JDF-D5); $\blacklozenge$, $1,500^\circ\text{C}$ (JDF-D4). Note that at our water concentration levels, $\bar{D}^*_{\text{water}}$ values calculated from equation (2) are almost independent of the choice of K because the dependences of $D_{\text{H}_2\text{O}}$ and $d[\text{H}_2\text{O}]/d[\text{water}]$ on K cancel: that is, at these low water contents, $D_{\text{H}_2\text{O}}$ is almost proportional to K , and $d[\text{H}_2\text{O}]/d[\text{water}]$ is almost inversely proportional to K (at very low water content⁷, $d[\text{H}_2\text{O}]/d[\text{water}] = 8[\text{water}]/K$). b, Apparent total water diffusivities at 0.2 wt% total water in a natural basaltic melt at 1.0 GPa (filled circles and upper solid line, this study) and in rhyolitic melt at 0.1 or 70 MPa extrapolated from lower temperature data^{4,7} (lower solid line). Nominal 1σ errors for diffusivities from this study are roughly the size of the symbol. Also shown are apparent carbonate diffusivities in an iron-free synthetic 'basaltic' melt at 1.5 GPa ($\circ$ and $---$)¹⁴ and in the JDF basalt at $1,300^\circ\text{C}$ and 1.0 GPa ($\diamond$, this study), and noble gas diffusivities in a tholeiitic basaltic melt at $1,350^\circ\text{C}$ and 0.1 MPa ($\blacklozenge$)³².

ution of bubbles. The profile can be fitted well if we assume a constant CO_3^{2-} concentration at the graphite-melt interface for the duration of the experiment and constant apparent carbonate diffusivity¹³ of $(13 \pm 1) \times 10^{-12} \text{ m}^2 \text{ s}^{-1}$ (nominal 1σ error). This diffusivity agrees well with the value ($\sim 11 \times 10^{-12} \text{ m}^2 \text{ s}^{-1}$) extrapolated from data¹⁴ for an iron-free model 'basaltic' melt at 1.5 GPa. Note that there must be a f_{O_2} profile accompanying the carbonate profile adjacent to the graphite²¹ in each of our experiments, and that determinations of carbonate concentrations in melts enclosed in graphite capsules can be used to measure f_{O_2} .

Diffusivity values at 0.2% water in basaltic melts calculated from equation (2) can be fitted by the following Arrhenius relation:

$$\ln D_{\text{water}}^* (\text{m}^2 \text{ s}^{-1}) = -(12.49 \pm 2.35) - (15200 \pm 3900)/T(\text{K}) \quad (3)$$

At other low water contents D_{water}^* can be estimated from the total water content, to which it is roughly proportional. Figure 2b compares diffusivities of various components. The activation energy for total water diffusion at 0.2% total water in the basaltic melt is $126 \pm 32 \text{ kJ mol}^{-1}$, slightly higher than that in rhyolitic melt ($\sim (103 - 23) = \sim 80 \text{ kJ mol}^{-1}$; see refs 4, 7 for details), but significantly lower than that for apparent carbonate diffusion in iron-free 'basaltic' melt ($249 \pm 30 \text{ kJ mol}^{-1}$; see ref. 14). Values of D_{water}^* at 0.2% water in the basaltic melt are 30–50 times those in rhyolitic melts and 7–10 times the apparent carbonate diffusivities in the JDF basalt or the iron-free 'basalt'. For the D_{water}^* values to be identical to the apparent carbonate diffusivity at 1,300 °C would require total water content as low as 0.02%, lower than the water contents of typical natural basaltic melts^{15,22–26}. Diffusivities of noble gas elements are in general higher than the apparent diffusivity of carbonate but may overlap with the total water diffusivity, depending on the water content.

These water diffusion data can be used to evaluate diffusive fractionation between different volatile components during bubble and crystal growth in basaltic melts. Watson *et al.*¹⁴ discussed such fractionations between carbon dioxide and water, but assumed that water diffusivities extrapolated from data in rhyolitic melt could be applied to basaltic melts; in other words, that the apparent diffusivity of total water in basaltic melts is lower than that of carbonate. Our data show that the opposite is true, and their conclusions must be reversed. For example, diffusive fractionation would increase the ratio of water to carbon dioxide in bubbles grown from oversaturated melt above that expected from equilibrium partitioning, but decrease the ratio in interface melts adjacent to advancing anhydrous phenocrysts.

Bubbles in mid-ocean-ridge basalts have varied ratios of $\text{H}_2\text{O}/\text{CO}_2$ and varied fractions of minor and trace components such as N_2 , H_2 , CO , SO_2 and noble gases^{27–29}. Part of these variations may be due to the diffusive control on bubble composition. Because mid-ocean-ridge basalts are often oversaturated with respect to CO_2 -rich vapour when they are erupted on the sea floor^{15,30} (that is, bubbles do not reach equilibrium with the bulk melt during rapid ascent), the diffusive transport of volatiles must play some part in controlling bubble growth rates and bubble composition. In addition, the compositions of multi-component gas bubbles (for example, the $\text{H}_2\text{O}/\text{CO}_2$, He/CO_2 ratios and noble gas patterns) in the transient bubble growth stage depend strongly on the solubilities and diffusivities of the various volatile components³¹. If equilibrium is not reached for any volatile component, each component would interact only by diffusive transport with the adjacent melt that is on average $\sim 2\sqrt{D_i t}$ away from the bubble, where D_i is diffusivity for component i . In this extreme case, bubbles could have a He/Xe ratio higher than expected from equilibrium partitioning. It is more likely that equilibrium can be reached for rapidly diffusing components such as He but not for slowly diffusing components such as CO_2 . Kinetic fractionations of this sort may be important

in the evolution of bubble compositions and compositional differences between smaller and larger bubbles. With our new data on water diffusivity in basaltic melts and previous data on carbonate and noble gas diffusivities^{14,32,33}, it should be possible to model in detail the evolution of bubble composition as a function of depth during magma ascent. It may also be possible to infer magma ascent rates from the diffusion profiles of volatile components adjacent to bubbles. □

Received 27 December 1990; accepted 10 April 1991.

1. Roberts, G. J. & Roberts, J. P. *Phys. Chem. Glass.* **5**, 26–32 (1964).
2. Burn, I. & Roberts, J. P. *Phys. Chem. Glass.* **11**, 106–114 (1970).
3. Delaney, J. R. & Karsten, J. L. *Earth planet. Sci. Lett.* **52**, 191–202 (1981).
4. Karsten, J. L., Holloway, J. R. & Delaney, J. R. *Earth planet. Sci. Lett.* **59**, 420–428 (1982).
5. Lapham, K. E., Holloway, J. R. & Delaney, J. R. *J. non-cryst. Solids* **67**, 179–191 (1984).
6. Stanton, T. R., Holloway, J. R., Hervig, R. L. & Stolper, E. M. *Eos* **66**, 1131 (1985).
7. Zhang, Y., Stolper, E. M. & Wasserburg, G. J. *Geochim. cosmochim. Acta* **55**, 441–456 (1991).
8. Doremus, R. H. in *Reactivity of Solids, Proc. 6th Int. Symp.* (ed. Mitchell, J. W.) 667–673 (Wiley Interscience, 1969).
9. Ernsberger, F. M. *Phys. Chem. Glass.* **21**, 146–149 (1980).
10. Smets, B. M. & Lommen, T. P. *Phys. Chem. Glass.* **24**, 635–636 (1983).
11. Wasserberg, G. J. *J. Geol.* **65**, 15–23 (1988).
12. Wakabayashi, H. & Tomozawa, M. *J. Am. Ceram. Soc.* **72**, 1850–1855 (1989).
13. Zhang, Y., Stolper, E. M. & Wasserburg, G. J. *Earth planet. Sci. Lett.* (in the press).
14. Watson, E. B., Sneeringer, M. A. & Ross, A. *Earth planet. Sci. Lett.* **61**, 346–358 (1982).
15. Dixon, J. E., Stolper, E. M. & Delaney, J. R. *Earth planet. Sci. Lett.* **90**, 87–104 (1988).
16. Zhang, Y., Walker, D. & Leshner, C. E. *Contr. Miner. Petrol.* **102**, 492–513 (1989).
17. Zhang, Y., Stolper, E. M. & Ihinger, P. D. V. M. *Goldschmidt Conf. Abstr.* 94 (1990).
18. Stolper, E. M. *Contr. Miner. Petrol.* **81**, 1–17 (1982).
19. Chekhmir, A. S., Epel'baum, M. B. & Simakin, A. G. *Geokhimiya* **10**, 303–305 (1988).
20. Crank, J. *The Mathematics of Diffusion* (Oxford University Press, Oxford, 1975).
21. Delaney, J. R., Muenow, D. W. & Graham, D. G. *Geochim. cosmochim. Acta* **42**, 581–594 (1978).
22. Muenow, D. W., Graham, D. G. & Liu, N. W. K. *Earth planet. Sci. Lett.* **42**, 71–76 (1979).
23. Harris, D. M. *J. Geol.* **89**, 689–701 (1981).
24. Harris, D. M. & Anderson, A. T. Jr. *Geochim. cosmochim. Acta* **47**, 1139–1150 (1983).
25. Kyser, T. K. & O'Neill, J. R. *Geochim. cosmochim. Acta* **48**, 2123–2133 (1984).
26. Moore, J. G., Batchelder, J. N. & Cunningham, C. G. *J. Volcan. geotherm. Res.* **2**, 309–327 (1977).
27. Pineau, F. & Javoy, M. *Earth planet. Sci. Lett.* **62**, 239–257 (1983).
28. Jambon, A. & Zimmermann, J. L. *Chem. Geol.* **62**, 177–189 (1987).
29. Stolper, E. M. & Holloway, J. R. *Earth planet. Sci. Lett.* **87**, 397–924 (1988).
30. Cable, M. & Frade, J. R. *J. Mater. Sci.* **22**, 919–924 (1987).
31. Lux, G. *Geochim. cosmochim. Acta* **51**, 1549–1560 (1987).
32. Carroll, M. R. *Eos* **71**, 1664 (1990).

ACKNOWLEDGEMENTS. We thank M. R. Carroll for comments and S. van der Laan for providing the JDF glasses; Y.Z. thanks M. B. Baker for technical assistance. This work is supported by NSF and NASA.

High-temperature electrical conductivity of the lower-mantle phase (Mg, Fe)O

B. J. Wood & J. Nell

Department of Geology, Bristol University, Wills Memorial Building, Queens Road, Bristol BS8 1RJ, UK

THE electrical conductivity of the lower mantle, which controls the transmission of geomagnetic signals to the Earth's surface, has been the subject of recent controversy^{1,2}. Peyronneau and Poirier¹ extrapolated high-pressure conductivity measurements of the lower-mantle assemblage (Mg, Fe)SiO₃ perovskite plus (Mg, Fe)O magnesiowüstite from 673 K to lower-mantle temperatures (2,200–2,500 K) to obtain conductivities consistent with geomagnetic observations ($\sim 1 \text{ S m}^{-1}$ at 1,000-km depth). The high-pressure study of Li and Jeanloz² gave conductivities several orders of magnitude lower. Here we report measurements of conductivity and thermopower of the iron-rich phase (Mg, Fe)O at high temperature (1,173–1,773 K) and controlled oxygen partial pressure (p_{O_2}). We find that p_{O_2} has a very large influence on conductivity: at fixed temperature and pressure, changes in p_{O_2} can vary the conductivity by 1.5 orders of magnitude within the stability field of magnesiowüstite. For plausible mantle values of p_{O_2} , temperature and bulk composition, we obtain results in good agreement with those of ref. 1, indicating that magnesiowüstite is the dominant conductor in the lower mantle and that observed upper- and lower-

mantle conductivities can be produced by the same bulk composition.

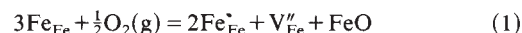
Magnesiowüstites with values of $\text{Mg}/(\text{Mg} + \text{Fe})$ of 0.0–0.98 were synthesized at 1,300 °C and 1 atm total pressure in a furnace with $p\text{O}_2$ controlled by flowing CO/CO_2 gas mixtures. After three cycles of grinding and firing, a slab ($\sim 12 \times 6 \times 1.5$ mm) was cut from a sintered pellet of the sample, its apparent density (0.8–0.9 times theoretical) measured, and the slab mounted on four type-B thermocouples. A small amount of platinum paint was applied to ensure good electrical contact. The sample was then placed in a 1-atm furnace and annealed at 1,400 °C for 12 hours before measurement at high temperature. This sample arrangement enabled us to make six measurements of thermopower between the six different thermocouple combinations. Conductivity (d.c.) was measured by passing a small current (10^{-4} A) between the Pt_{94} leads of the outer thermocouples and measuring the voltage drop across the inner thermocouple leads. The results were corrected for the small porosity and carefully monitored, at fixed temperature and $p\text{O}_2$, for any drift caused by loss of iron to the thermocouples. At high values of $\text{Mg}/(\text{Mg} + \text{Fe})$, a.c. conductivity was measured with electrodes attached to top and bottom of a disc of sample 9-mm diameter by 9-mm thick. A commercial a.c. source was employed and conductance measured in the frequency range 300–50,000 Hz using a Wayne-Kerr a.c. bridge.

Most of the measurements discussed here were made on samples with $\text{Mg}/(\text{Mg} + \text{Fe})$ of 0.83. This composition was

chosen as representative of an olivine-rich mantle with a bulk $\text{Mg}/(\text{Mg} + \text{Fe})$ ratio of 0.9. Under lower-mantle conditions iron partitions strongly into the magnesiowüstite phase³, which is expected to have $\text{Mg}/(\text{Mg} + \text{Fe})$ of 0.83 and to coexist with a $(\text{Mg}, \text{Fe})\text{SiO}_3$ perovskite having $\text{Mg}/(\text{Mg} + \text{Fe})$ of ~ 0.97 (refs 3 and 4). The conductivity of silicates and oxides in which iron is mainly divalent depends strongly on iron content^{5,6}, with iron end-members being many orders of magnitude more conducting than magnesium end-members. We therefore hypothesized that the magnesiowüstite phase, although volumetrically less important than perovskite, should be the dominant contributor to lower-mantle conductivity.

We find that, at each temperature, both electrical conductivity, σ , and thermopower, Q , depend strongly on the partial pressure of oxygen (Fig. 1). We observe that $\log_{10} \sigma$ is roughly proportional to $0.33 \log_{10} p\text{O}_2$, in agreement with earlier studies^{7,8} and that Q ranges from +350 to $-100 \mu\text{V K}^{-1}$ over the 1.5 log units of isothermal variation in σ . The effect of $p\text{O}_2$ on conductivity is extremely large; it is, for example, six times as large as the effect of raising temperature from 1,173 to 2,000 K at constant oxidation state (Fig. 2) and ~ 10 times the effect of raising the pressure by 10 GPa (refs 1, 9, 10). It seems, therefore, that oxidation state (uncontrolled in the high-pressure studies) is of prime importance in determining the properties of magnesiowüstite-bearing assemblages and that it is essential that it be controlled or measured. This is because magnesiowüstite, unlike most ferromagnesian silicates, can accommodate substantial concentrations of Fe^{3+} (refs 11, 12).

Conversion of Fe^{2+} in magnesiowüstite to Fe^{3+} with generation of a vacancy may be understood in terms of the equilibrium:



where Fe_{Fe} refers to 'normal' ferrous iron, $\text{Fe}_{\text{Fe}}^{\bullet}$ to ferric iron on the ferrous position, and $\text{V}_{\text{Fe}}^{\bullet}$ to a vacancy in the ferrous position. At low values of $p\text{O}_2$, $(\text{Mg}, \text{Fe})\text{O}$ of mantle-like composition is essentially ferric-free, whereas values of $\text{Fe}^{3+}/(\text{Fe}^{2+} + \text{Fe}^{3+})$ of up to 0.75 are stable at high $p\text{O}_2$ (ref. 12). Wüstite (FeO) and magnesiowüstites^{13,14} conduct by electron-hopping between Fe^{2+} and Fe^{3+} so that the highest conductivities are achieved when the donor and acceptor sites are present at comparable concentrations ($\text{Fe}^{3+}/(\text{Fe}^{2+} + \text{Fe}^{3+}) \approx 0.5$). This agrees with our observations (Fig. 1). The thermopower Q also depends on the $\text{Fe}^{3+}/\text{Fe}^{2+}$ ratio in small polaron conductors¹⁵.

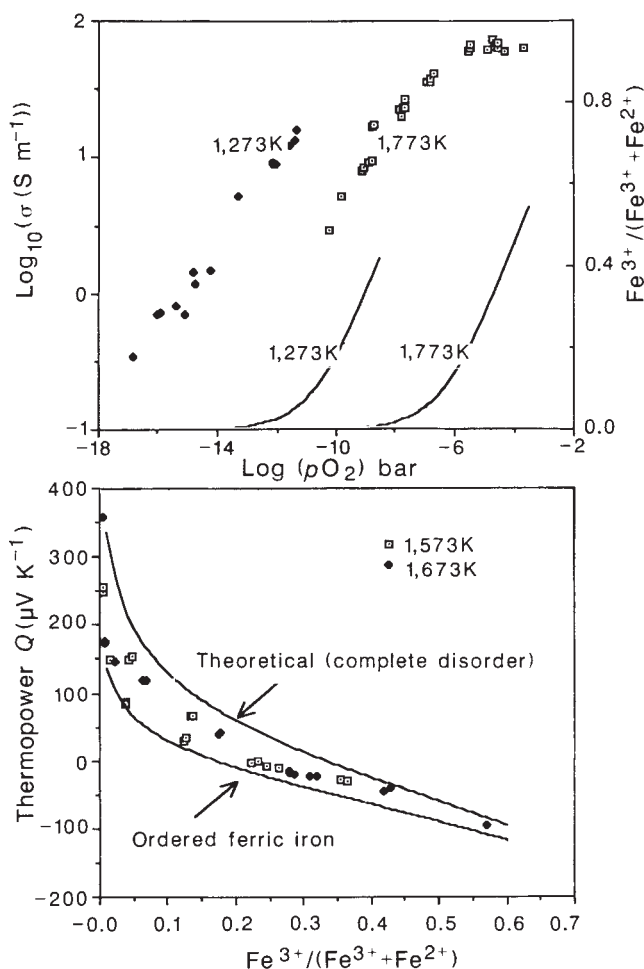


FIG. 1 Upper panel shows conductivity ($\log \sigma$) of $(\text{Mg}_{0.83}\text{Fe}_{0.17})\text{O}$ as a function of oxygen partial pressure at 1,273 and 1,773 K, and at 1 atm total pressure. Also shown is $\text{Fe}^{3+}/(\text{Fe}^{3+} + \text{Fe}^{2+})$ estimated from ref. 12. Lower panel shows thermopower, Q , as a function of Fe^{3+} content at 1,573 and 1,673 K. Curves are theoretical values calculated as discussed in the text.

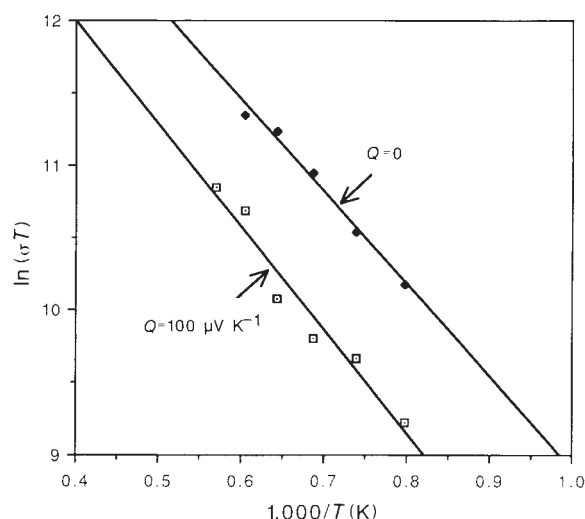


FIG. 2 Plot of $\ln \sigma T$ against reciprocal temperature at two values of thermopower Q . $Q=0$ corresponds to $p\text{O}_2$ close to the FMQ buffer; $Q=100 \mu\text{V K}^{-1}$ corresponds to $p\text{O}_2$ being 1–2 log units more reducing than FMQ.

If Fe^{3+} and Fe^{2+} were completely disordered in the structure, then for an n-type conductor, Q should be given¹⁵ by

$$Q = -\frac{k}{e} \left[\ln \left(\frac{2\text{Fe}^{3+}}{\text{Fe}^{2+}} \right) + A \right] \quad (2)$$

where k and e are Boltzmann's constant and the electronic charge, respectively, and A is an entropy of transport term which is usually small¹⁵. In Fig. 1 it can be seen that thermopower correlates in the expected way with the $\text{Fe}^{3+}/\text{Fe}^{2+}$ ratio, but that values are slightly lower than predicted from equation (2). Although this could be due to a non-zero value of A , it is more probably a result of $(\text{Fe}_{\text{Fe}} - V_{\text{Fe}}'' - \text{Fe}_{\text{Fe}})$ clustering, a phenomenon which is necessary to explain the observed $p\text{O}_2$ dependence of the $\text{Fe}^{3+}/\text{Fe}^{2+}$ ratio¹¹. Although the data do not allow us to determine the exact extent of clustering, we note that thermopower measurement is an *in situ* means of measuring the $\text{Fe}^{3+}/\text{Fe}^{2+}$ ratio. Thus, high-pressure experimentation, although not amenable to $p\text{O}_2$ control, could be accompanied by thermopower measurement, which would allow shifts in oxidation state to be established.

Extrapolation of our data to lower-mantle conditions requires a short extrapolation in temperature and a longer extrapolation in pressure at fixed oxidation state. The effect of raising the total pressure to a lower-mantle value of 30 GPa is about +0.6 log units in σ (refs 10, 17), substantially lower than the isothermal effects of $p\text{O}_2$. We have therefore ignored total pressure as a first approximation. The partial pressure of oxygen in the subcontinental upper mantle is close to the synthetic fayalite-magnetite-quartz (FMQ) buffer¹⁸, whereas oceanic regions seem to be ~ 1 log $p\text{O}_2$ unit more reducing¹⁸. The correlation between Q , Fe^{3+} content and $p\text{O}_2$ (Fig. 1) indicates that $Q=0$ roughly corresponds to FMQ conditions whereas values of +100 $\mu\text{V}/\text{K}$ correspond to 1–2 log $p\text{O}_2$ units below this buffer. Figure 2 shows the conductivities for these values of Q on a plot of $\ln \sigma T$ against $1/T$ (appropriate for small polaron conduction) which indicates that the activation energy for conduction is 0.55–0.61 eV. This corresponds to 0.42–0.48 eV in the more conventional interpretation of $\ln \sigma$ against $1/T$, a result in good agreement with earlier

workers^{7,8}. To compare our results with geomagnetic data and the high-pressure experiments^{1,2}, we need to know which phase, magnesiowüstite or perovskite, is the more conducting and what their relative proportions are in the lower mantle. Peyronneau and Poirier¹ found that perovskite and perovskite-magnesiowüstite mixtures of similar total $\text{Mg}/(\text{Mg} + \text{Fe})$ and oxidation state (nearly all iron present as Fe^{2+}) give similar conductivities at 40 GPa (Fig. 3). Because the Fe^{2+} partitions strongly into magnesiowüstite^{3,4} in the perovskite-magnesiowüstite mixtures of olivine $(\text{Mg}, \text{Fe})_2\text{SiO}_4$ stoichiometry used by these authors, the data require that $(\text{Mg}_{0.89}\text{Fe}_{0.11})\text{SiO}_3$ perovskite has essentially the same conductivity as a mixture of iron-poor $(\text{Mg}_{0.97}\text{Fe}_{0.03})\text{SiO}_3$ and $(\text{Mg}_{0.82}\text{Fe}_{0.18})\text{O}$. The only way this can occur is if the Fe-rich oxide phase has the main influence on overall conductivity in the olivine composition. The peridotitic upper mantle contains ~ 70 volume per cent $(\text{Mg}, \text{Fe})_2\text{SiO}_4$ olivine and when this disproportionates into perovskite plus $(\text{Mg}, \text{Fe})\text{O}$, about 22% of the latter phase would be produced. Numerical simulations of conductivity in isotropic mixtures¹⁹ yield overall values between 0.03 times (for $\sigma_{\text{cond}}/\sigma_{\text{resist}} = 10^5$, where σ_{cond} is the conductivity of the more conducting, σ_{resist} that of the more resistive component and 0.05 times ($\sigma_{\text{cond}}/\sigma_{\text{resist}} = 10^2$) the conductivities of the more conducting component. In either case the more conducting phase is the greater contributor to overall conductivity. Our data for $p\text{O}_2$ at FMQ and $p\text{O}_2$ at 1–2 log units below FMQ, when corrected for volume fraction of magnesiowüstite, agree well with the observed lower-mantle conductivity (Fig. 3).

In Fig. 3 we show the data of Peyronneau and Poirier¹, and Li and Jeanloz², for mixtures of magnesiowüstite with perovskite of overall composition $(\text{Mg}, \text{Fe})_2\text{SiO}_4$ (bulk $\text{Mg}/(\text{Mg} + \text{Fe}) \sim 0.9$), and recent high-pressure data on magnesiowüstites obtained by Li and Jeanloz¹⁰. The latter refer to samples synthesized under reducing conditions (>2 log units below FMQ) and have been corrected for dilution with a more resistive phase as described above. As can be seen, our data, those of Peyronneau and Poirier¹ and the high-pressure results on magnesiowüstite¹⁰ all agree with one another and indicate that magnesiowüstite is the important lower-mantle conductor. We therefore consider the Peyronneau and Poirier data¹ to be correct for the lower-mantle assemblage. The results also imply that the lower mantle has similar $p\text{O}_2$ and bulk chemical composition to the upper mantle.

The remaining question concerns the results of Li and Jeanloz² which are inconsistent with all the others. One possibility is that iron was lost to the internal platinum leads in these experiments, creating a resistive SiO_2 -rich layer, a phenomenon commonly observed in high-pressure experiments^{21,22}. It can, for example, be seen that their samples are even less conducting than pure MgO at the same temperature²⁰. Alternatively, the error may be due to large inhomogeneities in the laser-generated temperatures of their samples during the measurement¹⁷. □

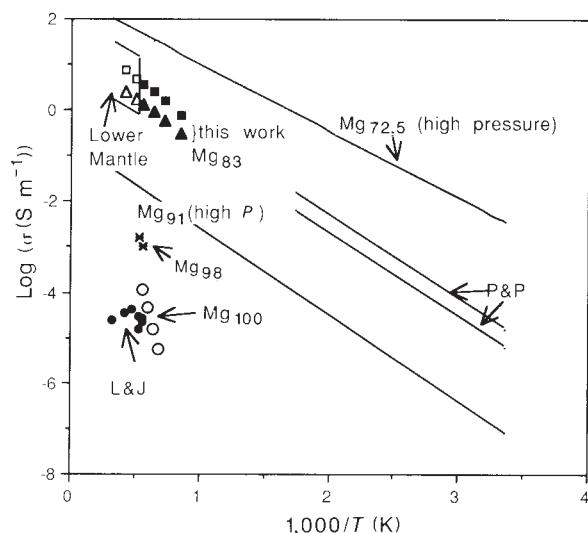


FIG. 3 Comparison of lower-mantle conductivity¹ with our results for $(\text{Mg}_{0.83}\text{Fe}_{0.17})\text{O}$, multiplied by 0.05 to take account of its volume fraction in the mantle. Squares refer to FMQ, triangles to 1–2 log units below FMQ. Open symbols are extrapolated. High-pressure data on $(\text{Mg}_{0.91}\text{Fe}_{0.09})\text{O}$ and $(\text{Mg}_{0.725}\text{Fe}_{0.275})\text{O}$, also multiplied by 0.05, are from ref. 10. The lines marked P & P refer to the range given by Peyronneau and Poirier¹ at 40 GPa for perovskite and perovskite-magnesiowüstite mixtures with $\text{Mg}/(\text{Mg} + \text{Fe})$ of 0.89. The data of Li and Jeanloz² (L & J) refer to a similar composition at simultaneous high pressure and temperature. Data for $(\text{Mg}_{0.98}\text{Fe}_{0.02})\text{O}$ are from this work. MgO results are from ref. 20.

Received 17 December 1990; accepted 2 April 1991.

- Peyronneau, J. & Poirier, J.-P. *Nature* **342**, 537–539 (1989).
- Li, X. & Jeanloz, R. *J. geophys. Res.* **95**, 5067–5078 (1990).
- Ito, E. & Takahashi, E. *J. geophys. Res.* **94**, 10637–10646 (1989).
- Wood, B. J. *J. geophys. Res.* **95**, 12681–12685 (1990).
- Hinze, E., Will, G. & Cernic, L. *Phys. Earth planet. Inter.* **25**, 245–254 (1981).
- Nell, J., Wood B. J. & Mason T. O. *Am. Miner.* **84**, 339–351 (1989).
- Iyengar, G. N. K. & Alcock, C. B. *Phil. Mag.* **21**, 293–304 (1970).
- Chen, W.-K. & Peterson, N. L. *J. Phys. Chem. Solids* **41**, 335–339 (1980).
- Mao, H. K. *Yb. Carnegie Inst. Wash.* **72**, 554–557 (1973).
- Li, X. & Jeanloz, R. *J. geophys. Res.* **95**, 21609–21612 (1990).
- Brynestad, J. & Flood, H. *Elektrochem. Ber. Bunsenges. Phys. Chem.* **62**, 953–958 (1958).
- Valet, P.-M., Pluschke, W. & Engell, H.-J. *Arch. Eisenhüttenwes.* **46**, 383–388 (1975).
- Gartstein, E., Cohen, J. B. & Mason, T. O. *J. Phys. Chem. Solids* **47**, 775–781 (1986).
- Hirsch, L. M. & Shankland, T. J. *J. geophys. Res.* **96**, 385–403 (1991).
- Chaikin, P. M. & Beni, G. *Phys. Rev.* **B13**, 647–651 (1976).
- Wu, C. C. & Mason, T. O. *J. Am. Ceram. Soc.* **64**, 520–522 (1981).
- Poirier, J.-P. & Peyronneau, J. *Proc. US–Japan High Pressure Miner. Phys. Symp.* (in the press).
- Wood, B. J., Bryndzia, L. T. & Johnson, K. E. *Science* **248**, 337–345 (1990).
- Webman, I., Jortner, J. & Cohen, M. H. *Phys. Rev.* **B14**, 4737–4740 (1976).
- Sempolinski, D. R. & Kingery, W. D. *J. Am. Ceram. Soc.* **63**, 664–669 (1980).
- Ito, K. & Kennedy, G. C. *Am. J. Sci.* **265**, 519–538 (1967).
- Nehru, C. E. & Wyllie, P. J. *J. Geol.* **83**, 455–471 (1975).

The relationship between abundance and body size in British birds

Sean Nee, Andrew F. Read, Jeremy J. D. Greenwood* & Paul H. Harvey

AFRC Unit of Ecology and Behaviour, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

* British Trust for Ornithology, Thetford, Norfolk IP24 2PU, UK

THE relationship between abundance and body size is the subject of considerable debate in ecology^{1–15}. Several data sets spanning a large range of body sizes show linear negative relationships between abundance and weight^{1–6} when these are measured on a logarithmic scale. But other studies of the abundances of species from single taxa, such as birds, which span a narrower range of body sizes reveal either little or no relationship, or a triangular relationship^{9–15}. Errors in estimating abundance might obscure relationships that do exist over a narrow range of body sizes. We describe here the relationship between body weight and abundance in British birds, whose population size estimates are unusually good. Abundance across all species declines with a -0.75 power of body weight, which conforms with the energetic equivalence 'rule'^{2,16}. There is, however, a significant positive relationship between abundance and body weight within lower taxa. Those tribes that do not share recent common ancestry with other British birds are most likely to show a positive relationship across their constituent species. We thus show that phylogenetic relatedness might be an important indicator of the structure of the relationship between body size and abundance.

As our measure of abundance, we used the numbers of breeding pairs published by the British Trust for Ornithology¹⁷ (BTO) for 147 species of British breeding birds. The BTO lists the abundances in Britain (estimated in a variety of ways¹⁷) of 228 species, from which we have eliminated occasional breeders, seabirds (which feed in a different ecosystem of undefined extent), species which have recently colonized Britain (such as the collared dove, *Streptopelia decaocto*), escapees and birds which are stocked (such as the red-legged partridge, *Alectoris rufa*). We included in the analysis only those species that have been regular annual breeders in Britain for the last few decades by consulting refs 17–19. Where an abundance range was given, we used the arithmetic mean. Body weights were obtained from refs 20 and 21. Bird body weights vary throughout the year so, where available, we used average adult female body weight in the spring and early summer. Otherwise, average female or average weight for both sexes was used.

The relationship between abundance and body weight in British birds (Fig. 1) arises in part because of a difference between passerines (typically abundant, small bodied species)

and all other birds (generally larger bodied and less abundant). Within passerines alone, and within non-passerines alone, there is no evidence of any association between population size and body weight ($r^2 = 0.002$, $n = 77$, $P = 0.69$; $r^2 = 0.007$, $n = 70$, $P = 0.48$, respectively). Even though non-passerines form a paraphyletic group, this suggests that the overall negative relationship, may be a higher-taxon phenomenon and may not hold within lower taxa.

We thus looked at the relationship within tribes. Taxonomy (and nomenclature) follows Sibley *et al.*²² for tribes and above and Howard and Moore²³ for species and genera. Sibley *et al.*²² use DNA hybridization data and we infer phylogeny from taxonomy. In 18 of the 25 tribes there is a positive relationship, more than expected by chance (binomial $P = 0.043$), and the z -transformed Spearman rank correlation coefficients are more positive than expected by chance ($t_{24} = 3.05$, $P = 0.006$). Thus, across closely related species, increases in body size are associated with increased population size, in contrast to the negative association across all species. This raises the question of what happens across more distantly related taxa.

Relationships across and within taxa can be analysed by considering correlations across species within genera, across generic means (calculated from constituent log-transformed species points) within tribes, across tribal means (calculated from constituent generic means) within subfamilies, and so on^{24,25}. Overall there is no consistent within-taxon association between abundance and body size (39 positive and 39 negative associations), but the relationships across higher taxa are opposite to those found within lower taxa (Table 1). Across species within genera, and genera within tribes, there are almost twice as many positive as negative associations between size and abundance (binomial $P = 0.072$). Across constituent taxa means within middle-ranked taxa (subfamilies, families and superfamilies), there are six positive and eleven negative associations (binomial $P = 0.33$). Across taxa within higher taxa (parvorder and above) there are twelve negative associations and four positive associations (binomial $P = 0.077$). The pattern within higher taxa is significantly different from that within lower taxa ($\chi^2_1 = 7.40$, $P = 0.007$). Thus there are positive associations between body weight and population size across closely related species and taxa, but when more distantly related taxa are compared, those characterized by larger bodied species have lower abundances. This pattern reversal is distinct from the often observed phenomenon of different allometric relationships between species as opposed to within them^{26,27}.

Damuth² has suggested as an empirical generalization that each species is likely to have the same amount of energy flowing through it, regardless of its body size. As metabolic rate increases with body weight to the 0.75 power²⁸, this ecological rule, if true, requires that species abundance decreases with body weight to the -0.75 power, as found for our data set (Fig. 1). But there is considerable scatter in the data ($r^2 = 0.14$) and the rule completely fails over relatively narrow but ecologically important

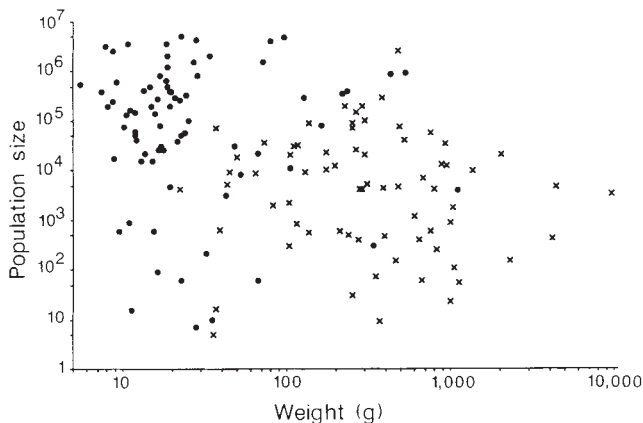


FIG. 1 The relationship between population size and body weight in British inland breeding bird species. ●, Passerines; ×, non-passerines. The slope of the least squares regression is -0.75 ($r^2 = 0.14$, $P < 0.0001$, $n = 147$). Model 1 regression is typically used to describe the relationship between body size and abundance (see for example ref. 2) and so we conform for the sake of comparability. The statistical error variance in body weight is small in relation to the variation among taxa in weight itself, thus justifying model 1 regression.

TABLE 1 Correlations between body weight and population size across subtaxa in different taxonomic levels

Across	Within	Number of positive correlations	Number of negative correlations
Species	Genera	16	9
Generic means	Tribes	13	7
Tribal means	Subfamilies	2	3
Subfamily means	Families	2	5
Family means	Superfamilies	2	3
Constituent higher taxa means*	Higher taxa	4	12
Total		39	39

* Across superfamilies within parvorders, parvorders within infraorders, infraorders within suborders, suborders within orders, orders within superorders, and across superorders within the class Aves.

size ranges such as that occupied by passerines. Furthermore, positive slopes are usually observed across closely related species. The cross-species negative relationship does not, therefore, adequately reflect the nature of the variation in the data. Note, however, that we observe the rule even though we are using a different measure of abundance from Damuth: he used 'ecological density', which is the population size in those areas where the species actually occurs, as opposed to population size in a specified geographical area.

Phylogeny is implicated in the relationship between body size and abundance in a very different way which may be relevant to understanding the positive relationship that exists at low taxonomic levels. We examined whether degree of relatedness of each tribe to other tribes in Britain has any effect on the relationship between body size and abundance within that tribe. Taxonomic levels were numbered 1 to 10 from the subfamily to the infraclass level. A tribe was denoted as joining the phylogenetic tree of British birds at a particular level on the criterion that it joins a taxon already containing more than half as many species as the original tribe. So, for example, the woodpecker tribe joins the tree at level 10 as it has no British sister taxon below the infraclass level, whereas the two warbler tribes, Phylloscopinae and Sylviini, join the tree at levels 2 and 3, respectively. Tribes with a positive relationship between abundance and body size are more likely to join the tree at higher taxonomic levels (Mann-Whitney $U = 30.5$, $P = 0.045$). That is, the more unrelated a tribe is to other tribes in Britain, the more likely it is to show a positive relationship. The six tribes that join the phylogenetic tree above the order level all have positive relationships.

Larger-bodied species may have an advantage in interspecific competition and the effects of such competition on resource access, either now or in the evolutionary past, may adversely affect the abundance of the smaller-bodied species. If so, we would be more likely to observe a positive relationship within tribes that encompass entire guilds. Phylogenetically distinct tribes, such as woodpeckers, owls and pigeons, probably include all the species having a distinct ecological role, whereas tribes with many close relatives, such as the warbler tribes, may not. Thus, insofar as phylogenetic divergence represents ecological divergence, tribes that are more likely to represent complete guilds are more likely to have a positive relationship.

The relationships described here, and the role of phylogeny in structuring these relationships, are so unexpected that we should determine whether they are a peculiarity of British birds. We thus examined this relationship in Swedish birds. Swedish bird abundance estimates may not be as good as the British ones owing to the inaccessibility of most of Sweden and the smaller number of ornithologists. The distinction between regular and irregular breeders may also be less clear on a subset (Sweden) of a much larger landmass, than on an island. Swedish

birds, however, do display many of the patterns we have described.

Swedish bird abundances were obtained from ref. 29, the body weights from refs 20 and 30 and those species were included that fit the same criteria as were used for the British birds. Across species of Swedish birds, the slope of the least squares regression of log abundance on log body weight is -0.77 ($r^2 = 0.18$, $P < 0.0001$, $n = 206$), conforming with the energetic equivalence hypothesis. Phylogenetic relatedness is also important; in addition to the lower taxa phenomena reported above for British birds, phylogeny is implicated in a novel way in the relationship between body size and abundance in Swedish birds. Ignoring, for the moment, the relationship across species within passerine genera, then within the remaining lower taxa (across generic means within all tribes and across species within non-passerine genera) there are almost twice as many positive as negative relationships (27 versus 15 taxa, binomial $P = 0.088$). The novel implication is that there are many more negative than positive relationships across species within genera of passerines (15 out of 17 genera, binomial $P = 0.002$). This is significantly different from the pattern within non-passerine genera ($\chi^2_1 = 11.43$, $P = 0.0007$), where there are almost twice as many positive as negative associations (15 out of 23 genera, binomial $P = 0.21$). Within both middle-ranked and higher taxa, negative relationships are no more common than expected by chance alone (11 out of 17 taxa, binomial $P = 0.33$; 8 out of 14 taxa, binomial $P = 0.79$, respectively). As in Britain, tribes with positive relationships are, on average, less related to other tribes present in Sweden, but in Sweden this pattern is not significant (Mann-Whitney $U = 80$, $P = 0.17$). Despite their greater potential inaccuracy, our analysis of the Swedish bird data also implicates phylogenetic relatedness as an indicator of the relationship between body size and abundance and further suggests that this indicator may reveal novel patterns in different regions. It will be interesting to see what the patterns are in other regions and in other groups of organisms. □

Received 5 December 1990; accepted 4 April 1991.

- Damuth, J. *Nature* **290**, 699–700 (1981).
- Damuth, J. *Biol. J. Linn. Soc., Lond.* **31**, 192–246 (1987).
- Peters, R. H. *The Ecological Implications of Body Size* (Cambridge University Press, 1983).
- Peters, R. G. & Wassenberg, K. *Oecologia* **60**, 89–90 (1983).
- Peters, R. H. & Raelson, J. V. *Am. Nat.* **124**, 498–517 (1984).
- Clutton-Brock, T. H. & Harvey, P. H. *J. Zool.* **183**, 1–39 (1977).
- Harvey, P. H. & Godfray, H. C. J. *Am. Nat.* **129**, 318–320 (1986).
- Sugihara, G. *Am. Nat.* **133**, 458–463 (1989).
- Brown, J. H. & Maurer, B. A. *Nature* **324**, 248–250 (1986).
- Juanes, F. *Am. Nat.* **128**, 921–929 (1986).
- Gaston, K. J. *Oikos* **53**, 49–57 (1988).
- Brown, J. H. & Maurer, B. A. *Am. Nat.* **130**, 1–17 (1986).
- Gaston, K. J. & Lawton, J. H. *Nature* **331**, 709–712 (1988).
- Morse, D. R., Stork, N. E. & Lawton, J. H. *Ecol. Entomol.* **13**, 25–37 (1988).
- Blackburn, T. M., Harvey, P. H. & Pagel, M. D. *J. Anim. Ecol.* **59**, 335–345 (1990).
- Odum, E. P. *Fundamentals of Ecology* 3rd edn (Saunders, Philadelphia, 1971).
- Marchant, J. H., Hudson, R., Carter, S. P. & Whittington, P. *Population Trends in British Breeding Birds* (British Trust for Ornithology, Tring, Herts, UK, 1990).
- Fisher, J. *South Eastern Naturalist and Antiquary* **5**, 1–10 (1952).
- Sharrock, J. T. R. *The Atlas of Breeding Birds in Britain and Ireland* (British Trust for Ornithology, Tring, Herts, UK, 1990).
- Brough, T. *Average Weights of Birds* (Ministry of Agriculture Fisheries and Food, Worplesdon Laboratory, Surrey, 1983).
- Cramp, S. & Simmons, K. E. L. *The Birds of the Western Palearctic* Vols 1–5 (Oxford University Press, 1977–1988).
- Sibley, C. G., Ahlquist, J. E. & Munroe, B. L. *Auk* **105**, 409–423 (1988).
- Howard, R. A. & Moore, A. *A Complete Checklist of Birds of the World* (Macmillan, London, 1984).
- Harvey, P. H., Clutton-Brock, T. H. & Mace, G. M. *Proc. natn. Acad. Sci. U.S.A.* **77**, 4387–4394 (1980).
- Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford University Press, 1991).
- Read, A. F. & Harvey, P. H. *J. Zool.* **219**, 329–353 (1989).
- Williamson, M. in *Biological Invasions: A Global Perspective* (eds Drake, J. A. et al.) 329–350 (Wiley, New York, 1989).
- Kleiber, M. *The Fire of Life* 2nd edn (Kreiger, New York, 1975).
- Ulfstrand, S. & Hogstedt, G. *Anser* **15**, 1–32 (1976).
- Perrins, C. *Birds of Britain and Europe* (Collins, London, 1987).

ACKNOWLEDGEMENTS. We thank T. Blackburn, S. Raivio, C. Godfray, J. Krebs, R. May, T. Guilford, A. Purvis and P. Cotgreave for helpful comments, the many volunteers who have collected the abundance data, R. Gregory for providing a checklist of Western Palearctic birds classified according to Sibley et al.²², S. Ulfstrand for bringing the Swedish data to our attention, and L. Gustafsson and T. Part for providing a list of Swedish species satisfying our criteria. S.N. is supported by the AFRC, A.F.R. by Christ Church, Oxford, UK.

Haploidy or diploidy: which is better?

Alexey S. Kondrashov* & James F. Crow†‡

* Research Computer Center, Academy of Sciences, Pushchino, Moscow Region, 142292, USSR

† Genetics Department, University of Wisconsin, Madison, Wisconsin 53706, USA

ALTHOUGH the evolutionary advantages of sexual reproduction have been extensively discussed¹⁻³, much less attention has been paid to haploid and diploid phases of the sexual life cycle. The relative lengths of these phases differ greatly in various taxa, including as extremes those with one or the other phase reduced to a single cell. Here we consider the efficiency of elimination of deleterious mutations as an evolutionary force and compare the mutation loads under haploid and diploid selection, L_n and L_{2n} . With truncation-like selection, partial dominance, and heterozygous effect of a mutation less than about 1/4 its hemizygous effect, $L_{2n} < L_n$; otherwise $L_{2n} > L_n$. The difference becomes important when the genomic deleterious mutation rate exceeds about 1 per genome. This suggests that the mutation rate, degree of dominance and mode of selection can be important in life-cycle evolution.

It has been suggested that selection in the diploid rather than the haploid phase increases fitness by reducing the effect of deleterious recessive and slightly dominant mutations. But this advantage is only transitory. Once equilibrium is reached, the diploid fitness is expected to be less than the haploid fitness because of the doubled mutation rate⁴.

Our key assumption is that the phase with the highest fitness (smallest mutation load) will increase in length, and we here identify a circumstance in which the diploid load is smaller than the haploid. The mutation load is $L = (w_{\max} - \bar{w})/w_{\max}$, where $\bar{w}$ is the mean fitness and $w_{\max}$ is the fitness of a (perhaps hypothetical) genotype with no mutations. For a single locus $L_n = \mu$, the per locus mutation rate, and $L_{2n} \approx 2\mu$, provided that selection is mainly in heterozygotes. For multiple, independently selected loci, $L_n \approx 1 - e^{-U}$ and $L_{2n} \approx 1 - e^{-2U}$, where $U = \sum \mu$, the haploid genomic mutation rate, so that $L_{2n} > L_n$ (refs 5, 6).

The mutation load can be reduced by synergistic epistasis (additional mutations having a greater deleterious effect than if they acted independently). For example, if the fitness of an individual with x equally deleterious mutations (expressed in heterozygotes or hemizygotes) is proportional to $1 - bx^2$, the load is reduced by about half⁷. The same is true if $w(x)$ is proportional to $\exp(-bx^2)$ (ref. 3). The diploid load is further reduced as the mutations become more nearly recessive, but it never becomes smaller than the haploid load⁶.

The situation can be quite different with the more extreme epistasis of truncation selection. Let $p(x)$ be the frequency of individuals with x mutations, and $\mu = \mu[p]$ and $\sigma^2 = \sigma^2[p]$ be the mean and variance of $p(x)$. We assume that all individuals with $x > T$ are eliminated, whereas the remainder are equally fit (Fig. 1). Let I be the selection differential, the difference in the mean number of mutations after and before selection. With a given shape of the distribution $p(x)$, I is determined by the proportion eliminated, which is the mutation load, L . At mutation-selection equilibrium, $-I = U$.

With heterozygous selection each mutation has a smaller effect on fitness than with hemizygous selection. If k heterozygous mutations are equivalent to a single hemizygous one, the fixed truncation point T will be proportional to x for haploids and to kx for diploids. Thus, measured in the number of mutations per individual, $T_{2n} = kT_n$. For what values of k is the diploid load less than the haploid?

To determine this, we first consider the case in which they are equal. If $p_n(x)$ and $p_{2n}(x)$ have the same shape and $L_n = L_{2n}$, the standardized selection differentials, I_n/σ_n and I_{2n}/σ_{2n} , will be the same. But to balance the doubled diploid mutation rate, I_{2n} must be twice I_n , so $\sigma_{2n} = 2\sigma_n$. If the ratio σ^2/μ is the same in haploids and diploids, $\mu_{2n} = 4\mu_n$.

We now justify the assumption that the equilibrium distribution shapes and variance/mean ratios remain the same in haploids and diploids. Numerical studies (A.S.K., unpublished) have shown that with free recombination the equilibrium distribution after reproduction but before selection, $p(x)$, is very close to normal, even when L is large enough to make the distribution after selection far from normal. With gametic ('linkage') equilibrium, $\sigma^2 = \mu$. Selection reduces σ^2 by an amount determined by L ; after meiosis and syngamy σ^2 returns half way to the linkage equilibrium value, assuming free recombination. Therefore, we expect σ^2/μ to equilibrate at a value lower than 1, and determined by L , which we are for the moment taking to be the same in haploid and diploid phases. Numerical studies show that σ^2/μ ranges from ~ 1.0 for weak selection, through 0.6 when $L = 1/2$, to 0.5 as $L \rightarrow 1$.

So, what is the ratio of $T_{2n}/T_n = k$, such that $L_{2n} = L_n$? Let t be the deviation of T from μ in units of σ . Equating $kT_n = k(\mu_n + t\sigma_n)$ to $T_{2n} = \mu_{2n} + t\sigma_{2n} = 4\mu_n + 2t\sigma_n$ yields $k = 4 - 2t\sigma_n/(\mu_n + t\sigma_n)$. Because t is of order unity and $\sigma_n < \sqrt{\mu_n}$, $k \approx 4$. When the truncation point for diploids is about 4 times that for haploids the loads are equal; if it is greater than this, the diploid load is smaller. With dosage compensation, the dominance h is equal to $1/k$.

Thus, when the dominance is less than about 1/4, diploidy is favoured, because the double number of mutations can be eliminated with a load smaller than in the haploid; if the dominance is greater than this, haploidy is favoured.

It is very unlikely that strict truncation occurs in nature. But something roughly approximating this is reasonable (Fig. 1). Such quasi-truncation is nearly as effective as truncation, provided that the width, $2d$, is not large^{8,9}. As with truncation selection, diploidy is favoured if h is less than about 1/4.

The less the dominance, the greater the load advantage of diploidy (Fig. 2). But our calculations assume only heterozygous selection. In the recessive limit the diploid and haploid loads are identical, so the curves in Fig. 2 must return to a value of 1 as $h \rightarrow 0$. There is strong evidence from *Drosophila*, however, that partial dominance for fitness is sufficiently great that homozygous mutations can be ignored¹⁰.

Going from haploidy to diploidy confers an immediate selective advantage by reducing the impact of partially dominant and recessive mutations. In some cases the initial advantage

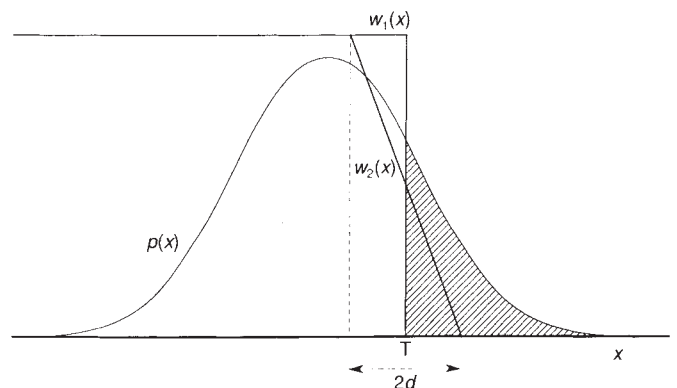


FIG. 1 Fitness functions under truncation selection, $w_1(x)$, and quasi-truncation $w_2(x)$, in which x is the number of deleterious mutations per individual and $p(x)$ is the frequency of x . The shaded area is the proportion saved with truncation selection and $2d$ is the width of the fitness-decreasing zone in standard deviation units.

‡ To whom correspondence should be addressed.

may be sufficient for stochastic fixation of diploidy. The initial advantage can lead to more than 99.99% diploids, equivalent to fixation in a small population, long before the diploid mutational equilibrium. Thus diploidy may become fixed even though $L_{2n} > L_n$. A return to haploidy is difficult because the mutational effects are no longer reduced by heterozygosity and the immediate deleterious consequence is equivalent to intense inbreeding in diploids⁴.

We have seen that with quasi-truncation, if h is less than about 1/4, diploidy will retain its load advantage at equilibrium. Thus there can be a systematic increase of the diploid phase and maintenance without stochastic fixation. We suggest, quite tentatively, that this may be one reason why diploidy has evolved. Our model for the efficacy of quasi-truncation selection in conferring an advantage on diploidy is quite crude. Yet, we believe that the simplifying assumptions are reasonable and provide a qualitative understanding.

Quasi-truncation selection is more likely to occur in density-regulated populations, whereas weak or no epistasis is more likely to be found in exponentially growing populations. This would suggest a correlation between the relative length of the two phases and the nature of selection, haploidy being favoured by r -selection and diploidy by K -selection.

There are other advantages of an extended diploid phase. One is protection from somatic mutations and chromosome loss, an argument first advanced by H. J. Muller⁴ and V. P. Efrimov¹¹. The redundancy of diploidy protects against the loss of function of essential cells and against recessive cancer mutations leading to pre-reproductive tumours such as retinoblastoma. Also, to whatever extent some overdominant loci are important, diploidy is favoured⁴.

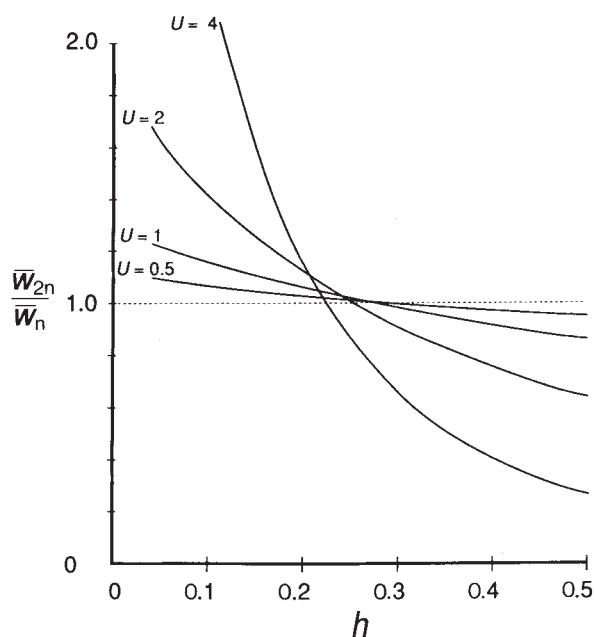


FIG. 2 The ratio of mean fitness of diploids to haploids, $\bar{W}_{2n}/\bar{W}_n$, as a function of the degree of dominance, h , of the mutations. The graphs are misleading in that they suggest that the ratio continues to rise as h becomes small. In reality, all the curves dip down to a value 1 as $h \rightarrow 0$: this is not shown because the exact shapes are not known. This downward slope is of little practical interest, however, because of the ubiquity of partial dominance. The values were calculated by computer iterations incorporating changes in mean and variance brought about by selection, mutation, meiosis with free recombination and syngamy. In all cases $T_n=10$ and $T_{2n}=10/h$. The four curves correspond to $U=0.5, 1, 2$, and 4 and the corresponding haploid loads are 0.111, 0.245, 0.505 and 0.836. Corresponding curves for quasitruncation with $d \leq \sigma$ are very similar and are not shown. Likewise, changes in T_n with corresponding changes in U make little difference in the curves or the critical dominance.

There is a well-known twofold cost associated with anisogamous sex. We have noted here an analogous twofold cost of diploidy because of the doubled mutation rate. Just as there must be compensatory advantages of sex if it is to prevail, there must be advantages of diploidy if it is to prevail. We have noted some ways in which this can happen.

Our comparison of loads is justified by the general increase of fitness by natural selection and by the fact that selection on modifiers usually goes in this direction¹². We might expect, however, that explicit models for the evolution of phase lengths will lead to qualitatively new results, including coexistence of selection at both phases.

Our arguments depend on the genomic deleterious mutation rate being large enough, of order 1 or higher, to produce a substantial mutation load. Although there is some support for such rates^{2,13}, the evidence is far from conclusive and experimental data are now needed. □

Received 7 January; accepted 12 March 1991.

1. Michod, R. E. & Levin, B. R. *The Evolution of Sex* (Sinauer, Sunderland, Massachusetts, 1988).
2. Kondrashov, A. S. *Nature* **334**, 435–440 (1988).
3. Charlesworth, B. *Genet. Res.* **55**, 199–221 (1990).
4. Crow, J. F. & Kimura, M. *Am. Nat.* **99**, 439–450 (1965).
5. Haldane, J. B. S. *Am. Nat.* **71**, 337–349 (1937).
6. Crow, J. F. in *Mathematical Topics in Population Genetics* (ed. Kojima, K. I.) 128–177 (Springer-Verlag, Berlin, 1970).
7. Kimura, M. *Jap. J. Genet.* **36S**, 179–190 (1961).
8. Milkman, R. *Genetics* **88**, 391–403 (1978).
9. Crow, J. F. & Kimura, M. *Proc. natn. Acad. Sci. U.S.A.* **76**, 396–399 (1979).
10. Crow, J. F. & Simmons, M. J. *The Genetics and Biology of Drosophila* Vol. 3c (ed. Ashburner, M., Carson, H. L. & Thompson, J. N.) 1–35 (Academic, London, 1983).
11. Efrimov, V. P. *J. de Biologie* **1**, 87–102 (1932, in Russian).
12. Karlin, S. & McGregor, J. *Theor. Pop. Biol.* **5**, 59–103 (1974).
13. Charlesworth, B., Charlesworth, D. & Morgan, M. T. *Nature* **347**, 380–382 (1990).

ACKNOWLEDGEMENTS. We thank Véronique Perrot, Andrei Skovoroda, Bill Engels and Brian Charlesworth for assistance and advice.

Transition from haploidy to diploidy

Véronique Perrot*†‡, Sophie Richerd* & Myriam Valéro*

* Laboratoire de Génétique et Evolution des Populations Végétales, URA CNRS 1185, Bât. SN2, USTLFA, F-59 655 Villeneuve d'Ascq cedex, France
† Zoologisches Institut der Universität, Rheinsprung 9, CH-4051 Basel, Switzerland

AS a direct consequence of sex, organisms undergo a haploid and a diploid stage during their life cycle. Although the relative duration of haploid and diploid phases varies greatly among taxa, the diploid phase is more conspicuous in all higher organisms. Therefore it is widely believed that diploidy offers more evolutionary possibilities^{1–3} and is thus nearly always selected for. We have now performed computer simulations to investigate one possible advantage of diploidy, that is, protection against the expression of deleterious mutations. Instead of comparing isolated haploid and diploid populations, we considered interbreeding haploids and diploids. Diploids invaded the population only when the dominance degree of a single deleterious mutation was smaller than about 1/2, and the condition allowing diploidy to invade depended on how harmful the mutation was.

The classical explanation for the success of diploidy relies on the protection it offers against expression of deleterious mutations⁴. Consider the simple case of one deleterious allele in a large population. The deleterious allele appears in each generation by mutation, but is selected against, and reaches a stable frequency balanced between mutation and selection⁵. Because this allele is deleterious, it reduces the mean fitness of the

‡ Present address: Zoologisches Institut der Universität, Rheinsprung 9, CH-4051 Basel, Switzerland.

TABLE 1 Differential survival of adults according to genotype at their fitness locus

	Genotype	Frequency	Fitness
Haploids	<i>A</i>	p_n	1
	<i>a</i>	q_n	$1-s$
Diploids	<i>AA</i>	p_{2n}^2	1
	<i>Aa</i>	$2p_{2n}q_{2n}$	$1-hs$
	<i>aa</i>	q_{2n}^2	$1-s$

The allele *A* mutates in *a* at the per generation mutation rate u . s , Selection coefficient; h , dominance degree; p_n and q_n (respectively p_{2n} and q_{2n}), frequencies of *A* and *a* in the haploid compartment (respectively in the diploid compartment). The fitness locus segregates independently of the cycle locus.

population. If part of a haploid population in mutation-selection balance becomes diploid, these diploid individuals would have higher mean fitness than the rest of the haploid population in the first generation if the deleterious allele is more or less recessive (dominance degree of the deleterious allele $h < 1/2$) (Fig. 1a, traces 1 and 3). But once diploids reach their own mutation-selection balance, the diploid population suffers larger fitness reduction than the haploid, unless the allele is completely recessive ($h = 0$) (Fig. 1a, traces 1 and 2). With a single deleterious mutation, the initial advantage of diploids is too small for them to become fixed before they reach their equilibrium. Therefore, in competition with haploids, diploids cannot persist at equilibrium. When haploids and diploids do not interbreed with each other, as might happen if they produce gametes at different times, it seems that the sheltering of deleterious mutations cannot account for the evolution of diploidy.

There are two ways in which this model can be modified. One can consider more mutations having nonadditive effects on fitness. For example, when the number of deleterious alleles per genome is randomly distributed in the population, and

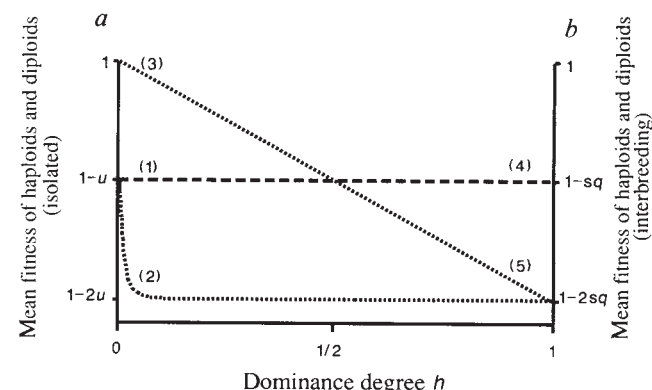


FIG. 1 Comparison of the mean fitness reduction due to a deleterious allele in haploids and diploids. u , Mutation rate; s , selection coefficient. *a*, Haploids and diploids do not interbreed. The mean fitness of isolated population is plotted against the dominance degree of the deleterious allele h ; the y axis is to be read on the left-hand side. Trace 1, In a haploid population at equilibrium, the mean fitness always equals $1-u$. Trace 2, In a diploid population at equilibrium, the mean fitness is $1-2u$, unless h is very small¹⁴. Trace 3, At the first generation of transition from haploidy to diploidy, that is when the deleterious allele frequency in the diploid population q_{2n} equals the equilibrium frequency in the haploid population, the mean fitness of the diploid population is $1-2uh$, provided that $u \ll s$. When the diploid population evolves towards its own equilibrium, q_{2n} increases, and therefore the diploid mean fitness decreases. *b*, Haploids and diploids do interbreed. The mean fitness of the haploid and the diploid group is plotted against h ; the y axis is to be read on the right-hand side. q , Frequency of the deleterious allele in the population. Trace 4, the mean fitness of the haploid group is $1-sq$. Trace 5, in first approximation, the mean fitness of the diploid group is $1-2hsq$.

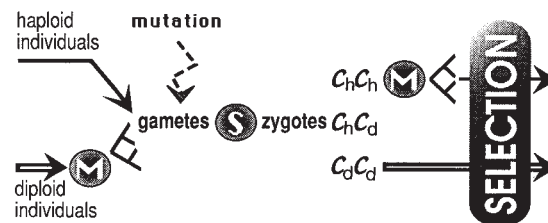


FIG. 2 Polymorphism for the life cycle: meiosis (M) occurs either just before gamete production (diploid phase expanded, haploid phase reduced to gametes: lower part of the scheme) or just after syngamy S (diploid phase reduced to the zygote, haploid phase expanded: upper part of the scheme). The fate of a zygote is determined by its own genotype at the cycle locus *C*. This locus has two alleles: C_h , the 'haploid' allele, which causes meiosis just after syngamy, and C_d , the 'diploid' allele, which delays meiosis until gamete production. Thus, a zygote can be of three possible genotypes at this locus: $C_h C_h$, $C_h C_d$, or $C_d C_d$. There are three possible genotypes for the individuals: if C_h is dominant, these are C_h and C_d (haploid individuals), and $C_d C_d$ (diploid individuals); if C_d is dominant, these are C_h (haploid individuals), and $C_h C_d$ and $C_d C_d$ (diploid individuals). Gametes are produced at the same time and mate at random. Mutation occurs at the gamete stage. Haploid and diploid reproductive cycles are selectively neutral.

individuals bearing more deleterious alleles than a threshold value have zero fitness, the diploid population suffers less fitness reduction than the haploid for $h < 1/4$ (ref. 6). We take a different and novel approach in comparing the dynamics of interbreeding haploid and diploid individuals, that is, a population polymorphic for the life cycle. In other words, instead of group-level selection, we consider individual-level selection.

We investigated a population containing both haploid and diploid individuals. Both types produced their gametes at the same time and mated at random with each other (Fig. 2). Into such a population, we introduced a deleterious allele, appearing by mutation in gametes, and being selected against at the adult stage, haploid or diploid respectively (Table 1). Only one type of individual persisted at equilibrium. Diploids invaded when h was $\leq 1/2$; haploids invaded when h was $\geq 1/2$.

Because haploids and diploids interbreed, and because the fitness and the cycle loci segregated independently, the genes selected in the haploid or diploid state were mixed in each generation. Consequently, the frequency of the deleterious allele was about the same in haploid and diploid individuals (q in Fig. 1b). Mean fitness of the haploid group was $1-sq$ (Fig. 1b, trace 4), and mean fitness of the diploid group was roughly $1-2hsq$ (Fig. 1b, trace 5). Thus, for a more or less recessive allele ($h < 1/2$), diploids had a higher mean fitness than haploids (Fig. 1b, traces 4 and 5, from left-hand side to $h = 1/2$) and diploids invaded the population. With $h > 1/2$, it was the other way around (Fig. 1b, traces 4 and 5, from right-hand side to $h = 1/2$), and haploids invaded the population.

The threshold value for h was not exactly $1/2$ (Fig. 3). Because

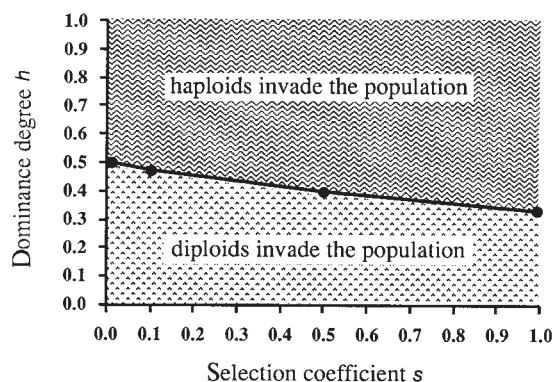


FIG. 3 Threshold value of the dominance degree h as a function of the selection coefficient s .

of different selection pressures on haploid and diploid individuals, the allele *A* at the fitness locus was more often associated with the 'haploid' allele than with the 'diploid' allele at the cycle locus in gametes, and the allele *a* was more often associated with the 'diploid' allele than with the 'haploid' allele (V.P., manuscript in preparation). This linkage disequilibrium caused the threshold value for *h* to be smaller than 1/2. For almost neutral mutation ($s \approx 0$), the threshold value for *h* was almost 1/2. The linkage disequilibrium was stronger when the mutation was more deleterious (*s* larger), and therefore the deviation from 1/2 was larger (V.P., manuscript in preparation). The deviation was identical whatever the dominant allele at the cycle locus and the mutation rate *u*.

If transition towards diploidy took place as supposed in this model, we have to reconsider classical theories of dominance. Fisher believed that the quasi-general recessiveness of deleterious alleles results from selection for modifier loci chang-

ing intermediate dominance into recessiveness⁷. To work, this theory requires diploidy to evolve first, because our results indicate that intermediate dominance does not allow diploids to invade, especially for strongly deleterious mutations. On the contrary, Wright's theory of dominance⁸, which considers deleterious mutations to be intrinsically more or less recessive, complies with our results. This complements other evidence in favour of Wright's theory^{9,10}.

In previous discussions about the evolution of haploid and diploid life cycles, authors have concluded that the diploid life cycle is advantageous^{4,11,12}. Our results delimit the conditions under which diploidy will invade when haploids and diploids interbreed. The fitness reduction suffered by the population once equilibrium is reached is independent of the harmfulness of a mutation¹³. However, the realm for diploidy depends not only on how dominant the mutant allele is, but also how harmful it is. □

Received 18 March; accepted 22 April 1991.

1. Stebbins, G. L. *Variation and Evolution in Plants* (Columbia University Press, New York, 1950).
2. Maynard Smith, J. *The Evolution of Sex* (Cambridge University Press, Cambridge, 1978).
3. Bernstein, H., Byerly, H. C. & Michod, R. E. *Am. Nat.* **117**, 537-549 (1981).
4. Crow, J. F. & Kimura, M. *Am. Nat.* **99**, 439-450 (1965).
5. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
6. Kondrashov, A. S. & Crow, J. F. *Nature* **351**, 314-315 (1991).
7. Fisher, R. A. *Am. Nat.* **62**, 115-126 (1928).
8. Wright, S. *Am. Nat.* **68**, 25-53 (1934).

9. Charlesworth, B. *Nature* **278**, 848-849 (1979).
10. Kacser, H. & Burns, J. A. *Genetics* **97**, 639-666 (1981).
11. Lewis, J. & Wolpert, L. *J. theor. Biol.* **78**, 425-438 (1979).
12. Paquin, C. & Adams, J. *Nature* **302**, 495-500 (1983).
13. Haldane, J. B. S. *Am. Nat.* **71**, 337-349 (1937).
14. Kimura, M. *Jap. J. Genet. Suppl.* **36**, 179-190 (1961).

ACKNOWLEDGEMENTS. We thank J. Crow, A. Kondrashov, J. Cugen, R. Hoekstra, B. Clarke, B. Olle Bengtsson, S. Stearns, A. van Noordwijk, J. Shykoff, R. Nager, C. Müller, F. Vollrath, D. Couvet, F. Kjellberg and P.-H. Gouyon for discussion.

HBx gene of hepatitis B virus induces liver cancer in transgenic mice

Chang-Min Kim*, Kazuhiko Koike†, Izumu Saito‡, Tatsuo Miyamura‡ & Gilbert Jay*§

* Laboratory of Virology, Jerome H. Holland Laboratory, American Red Cross, Rockville, Maryland 20855, USA

‡ Laboratory of Hepatitis Viruses, Department of Enteroviruses, National Institute of Health, Tokyo 141, Japan

THE exact role of hepatitis B virus in the development of liver cancer is not known. The recent identification of a viral regulatory gene *HBx* suggests a possible direct involvement of the virus whereby the *HBx* protein, acting as a transcriptional transactivator of viral genes, may alter host gene expression and lead to the development of hepatocellular carcinoma. We have tested this possibility by placing the entire *HBx* gene under its own regulatory elements directly into the germline of mice. Transgenic animals harbouring this viral gene succumbed to progressive histopathological changes specifically in the liver, beginning with multifocal areas of altered hepatocytes, followed by the appearance of benign adenomas, and proceeding to the development of malignant carcinomas. Male mice developed disease and died much earlier than females. This transgenic animal model appears ideal for defining the molecular events that follow the expression of the viral *HBx* gene and are responsible for the development of liver cancer.

Despite overwhelming epidemiological evidence linking hepatitis B virus (HBV) to the development of hepatocellular carcinoma¹⁻³, there are few experimental data to implicate this virus directly in liver malignancy^{4,5}. Indeed, most studies point either to the requirement of a second genetic event that results from continuous cell death and regeneration characteristic of chronic hepatitis^{6,7}, or to the activation by the HBV transcriptional control element of a cellular gene at the site of viral

integration⁸⁻¹⁰. An alternative mechanism of liver carcinogenesis directly involves viral gene products¹¹; this is particularly attractive in view of the more recent finding that HBV encodes a viral transactivator, the *HBx* antigen, which at least in tissue culture cells can function as a transcriptional control element to upregulate the expression of other viral genes¹²⁻¹⁴. As viral transactivators act through the host biosynthetic machinery, they should perturb the expression of cellular genes and the differentiated functions of the infected cell¹⁵⁻²⁰. We have investigated whether the expression of the *HBx* gene alone can induce liver cancer in transgenic mice.

The *HBx* transgenic mice were derived by microinjection of a 1.15-kilobase (kb) HBV subtype *adr* DNA fragment, which spans nucleotide positions 707 to 1,856 in the viral genome²¹, into single-cell embryos derived from outbred CD1 mice¹⁵⁻²⁰. This segment of DNA contains not only the entire coding region of the *HBx* gene (map positions 1,246 to 1,710), but also the transcriptional enhancer (map positions 945 to 1,106; ref. 22), the principal RNA start sites (including positions 1,184, 1,203 and 1,214; ref. 23), and the polyadenylation site (position 1,788; ref. 21). Six transgenic mice, each with at least one intact copy of the transgene stably integrated into the host genome, were identified by Southern blot analysis, and three of them (designated C11, H9 and E1) were randomly selected to be bred out into permanent lines.

Expression of the *HBx* transgene in different tissues was determined by northern blot analysis. Relatively high levels of expression of the 0.7-kb transcript²⁴ were observed in liver, kidney and testis (Fig. 1a). Similar results were obtained for all three transgenic lines. Selective expression in liver and kidney is consistent with the detection of HBV DNA sequences in the same tissues of infected individuals²⁵. That the *HBx* gene is expressed in the liver of the transgenic mice is confirmed by immunoprecipitation of liver extracts using a rabbit polyclonal antibody against *HBx*. A component of relative molecular mass 17,000 befitting the *HBx* protein²⁶ was specifically detected with the immune serum (Fig. 1b).

At the age of about 4 months, histological examination of the liver revealed the presence of multifocal areas of altered hepatocytes in all three strains of transgenic mice (Fig. 2a) but not in their normal littermates. These altered foci were made

† Present address: First Department of Internal Medicine, University of Tokyo, Tokyo 113, Japan.

§ To whom correspondence should be addressed.

up of cells with a poorly stained cytoplasm which is very granular, but show no evidence of increased cell proliferation. Consistent with their being preneoplastic lesions²⁷, there is cytoplasmic accumulation of glycogen detectable by periodic acid-Schiff staining (Fig. 2*b*). Cells within these altered foci also expressed a relatively high level of the HBx protein, as revealed by immunostaining (Fig. 2*c, d*), thus associating the histological change with the expression of the transgene in the same subset of hepatocytes.

At about 8 to 10 months of age, tumour nodules began to appear in the liver of the transgenic mice. These nodules were frequently made up of basophilic cells which proliferate to compress the neighbouring normal liver parenchyma (Fig. 3*a, b*) and have all the characteristics of hepatocellular adenomas²⁸. Like the cells in the altered foci, those that made up the nodules also accumulated relatively high amounts of the HBx protein (Fig. 3*a*) and contained increased levels of glycogen (Fig. 3*b*). But unlike in the altered foci, α -fetoprotein is detected in the adenomas (data not shown). Not all adenomas were of the basophilic type. Some made up of clear cells were also detected²⁹. On occasion, the clear cells seemed to arise from within a tumour nodule composed of basophilic cells (Fig. 3*c*). These tumour cells continued to express the HBx protein (Fig. 3*e*), and also to accumulate glycogen (Fig. 3*d*) and α -fetoprotein (Fig. 3*f*). Serum alanine aminotransferase concentration was consistently within the normal range in transgenic mice that after killing were found to have one or more adenomas in the liver. This suggests that liver cell damage is not associated with the development of these benign tumour lesions⁷.

Most of the male transgenic mice of the C11 strain died between 11 and 15 months of age. On autopsy, they all had liver tumours which were grossly detectable. Many of these lesions were diagnosed as hepatocellular carcinoma of the clear cell type^{30,31}. The extent of the lesions varied from the involvement of every single lobe of the liver to focal malignant areas overlying

TABLE 1 Susceptibility of transgenic mice to the development of liver tumours

Transgenic strain	Sex	No. of mice observed	No. of mice with tumours	Per cent of mice with tumours
C11	M	21	19	90.5
	F	20	12	60.0
	total	41	31	75.6
H9	M	10	8	80.0
	F	6	4	66.7
	total	16	12	75.0

a benign adenoma that affected an entire hepatic lobe (Fig. 3*g*). The clear cell carcinomas were made up of cells that were frequently binucleated. Like adenoma cells, they continue to express HBx antigen (Fig. 3*i*) and α -fetoprotein (data not shown). Unlike the benign cells, the carcinoma no longer had any accumulated glycogen (Fig. 3*h*). Thus, the apparent morphological progression was accompanied by biochemical change. Metastases were detected in some of the animals. It is important to note that control male CD1 mice have a lifespan of about 24 months, and fewer than 10 per cent showed any sign of hepatic neoplasms at the time of death, and none of these were of the clear cell type.

The penetrance of hepatocellular carcinoma in the C11 transgenic line is about 90 per cent for males and 60 per cent for females (Table 1). This observation appears to be consistent with epidemiological data suggesting that men chronically infected with HBV are much more prone to liver cancer than women². But if the age at which the mice develop the disease is taken into account, most males died with hepatocellular carcinoma at between 11 and 15 months, and the females at between 17 and 21 months (Fig. 4). Given that control CD1 mice normally die at 21 to 24 months of age, we suspect that with a longer lifespan the penetrance of hepatocellular carcinoma among females would be as high as among males; the difference between the sexes would seem to be when the disease develops, with both sexes being equally susceptible.

It is important when studying disease phenotypes in transgenic mice to avoid possible integration site artefacts, and this is achieved by analysing multiple founder mouse lines. All three transgenic strains developed hepatocellular carcinoma selectively. A comparison between the C11 and H9 transgenic lines showed no significant difference in the incidence of disease in male or female mice. Despite the expression of the *HBx* gene in the kidney and the testis, no histological change was observed in either tissue.

Although transgenic mice containing different HBV genes, including the entire viral genome, have been analysed before³²⁻³⁶, there has been little evidence to suggest that the virus has a direct role in inducing liver cancer. In those constructs where the *HBx* gene was included, its expression was not detected³³⁻³⁷. When the *HBx* gene was introduced under the transcriptional control of the human α -1-antitrypsin promoter, no liver tumours were found³⁷; as expression of the HBx antigen in these mice was shut off within 4 weeks after birth, progressive neoplastic changes in the liver would not be expected.

Our observation demonstrates that expression of the *HBx* gene alone will result in progressive morphological and, presumably, biochemical changes leading ultimately to the development of malignant liver cancer. Interestingly, even the HBx antigen is not expressed in all hepatocytes, but is restricted to a select subset, most probably those at a specific differentiation state, resulting in multifocal rather than generalized disease in the liver.

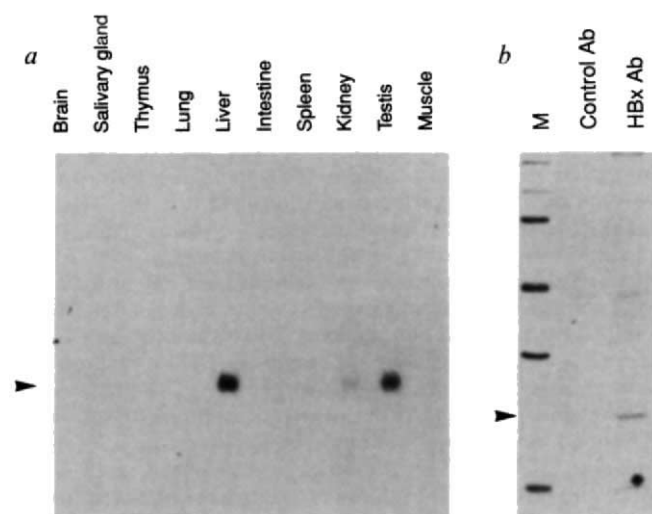


FIG. 1 Analysis of the expression of the *HBx* gene in transgenic mice. *a*, Northern blot hybridization analysis using the *HBx* DNA probe of poly(A)⁺ RNA extracted from tissues of a 2-month-old H9 transgenic mouse³⁸. Arrow indicates the 0.7-kb *HBx* transcript. *b*, SDS-polyacrylamide gel electrophoresis analysis of immunoprecipitates, obtained with either a control rabbit serum or a rabbit anti-HBx serum directed against a C-terminal peptide (amino acids 139 to 154), from a liver extract of a 3-month-old C11 transgenic mouse which had been injected intraperitoneally with 1 mCi each of ³⁵S-methionine and ³⁵S-cysteine³⁹. Arrow indicates the HBx protein of *M_r* 17,000 (17 K). Molecular weight markers (lane M) were 92.5, 68, 45, 30 and 12.5K. Ab, antibody.

FIG. 2 Early focal lesions of altered hepatocytes in the liver of transgenic mice. *a* and *b*, Serial sections of liver from a C11 transgenic male mouse killed at 6 months and stained with haematoxylin and eosin, or periodic acid-Schiff, respectively. *c* and *d*, Serial sections of liver from a C11 transgenic male mouse killed at 5 months, immunostained with either a rabbit anti-HBx serum directed against a C-terminal peptide (amino acids 139 to 154) or a control rabbit serum, respectively, and visualized by the avidin-biotin complex method (Vector Lab.). Arrows indicate foci of altered hepatocytes.

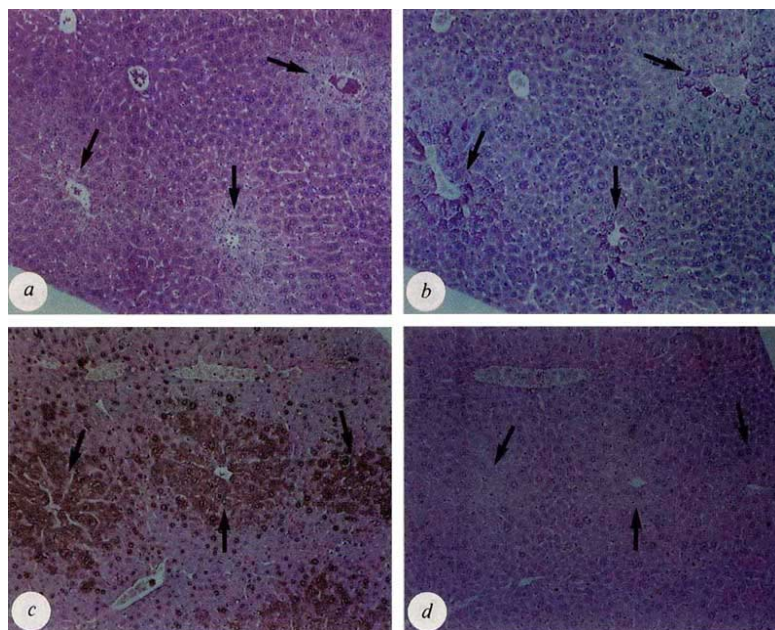
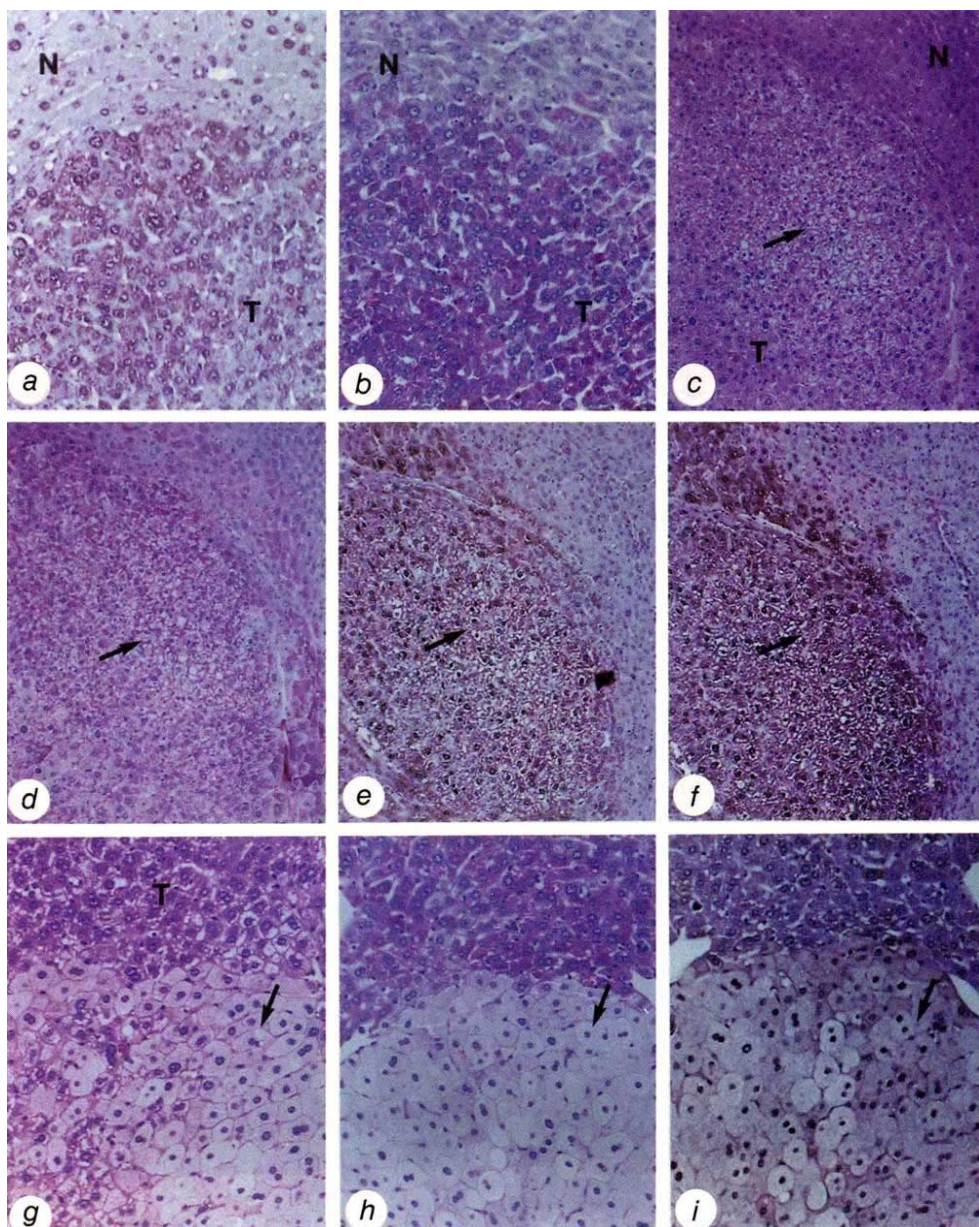


FIG. 3 Progressive neoplastic lesions in the liver of transgenic mice. *a* and *b*, Serial sections of liver from a 9-month-old C11 transgenic mouse, stained with a rabbit anti-HBx serum and periodic acid-Schiff, respectively. The tumour nodule (T) and the adjacent normal part of the liver (N) are indicated. *c-f*, Serial sections of liver from a 10-month-old H9 transgenic mouse stained with haematoxylin and eosin, periodic acid-Schiff, rabbit anti-HBx serum, and rabbit anti- α -fetoprotein, respectively. Arrow indicates the clear cell adenoma. *g-i*, Serial sections of liver of a 16-month-old E1 transgenic mouse stained with haematoxylin and eosin, periodic acid-Schiff, and rabbit anti-HBx serum, respectively. Arrow indicates the clear cell carcinoma.



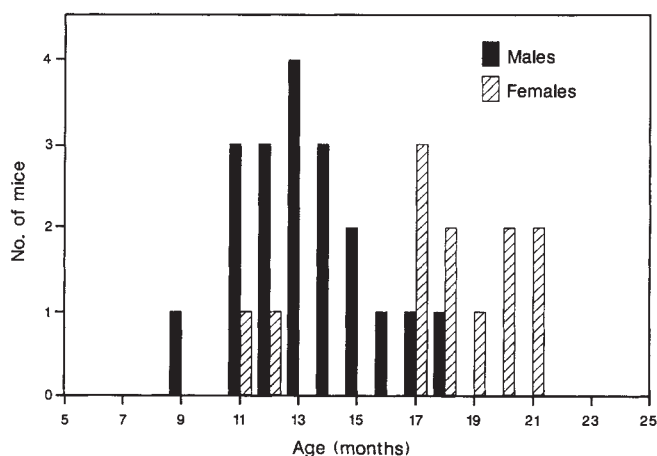


FIG. 4 Age at the time of death of transgenic mice with hepatic tumours. Transgenic C11 mice that died were subjected to gross and microscopic examination. Nineteen of 21 males and 12 of 20 females had detectable tumours in the liver.

In conclusion, our findings indicate that HBV is directly involved in the development of liver cancer and argue against chronic hepatitis being an absolute prerequisite for the aetiology of hepatocellular carcinoma in man. Also, the transgenic mice described here serve as a model system for defining the molecular events after the expression of the *HBx* gene that are responsible for the progressive changes in the liver. □

Received 21 November 1990; accepted 16 April 1991.

1. Tiollais, P., Pourcel, C. & Dejean, A. *Nature* **317**, 489–495 (1985).
2. Beasley, R. B. & Hwang, L.-Y. in *Viral Hepatitis and Liver Disease* (eds Vyas, G. N.) 209–224 (Grune & Stratton, Florida, 1984).
3. Beasley, R. P., Hwang, L.-Y., Linn, C. C. & Chien, C. S. *Lancet* **ii**, 1129–1133 (1981).
4. Dejean, A., Sonigo, P., Wain-Hobson, S. & Tiollais, P. *Proc. natn. Acad. Sci. U.S.A.* **81**, 5350–5354 (1984).
5. Rogier, C. E. *et al. Science* **230**, 319–322 (1985).
6. Chisari, F. V. *et al. J. Virol.* **60**, 880–887 (1986).
7. Chisari, F. V. *et al. Cell* **59**, 1145–1156 (1989).
8. de The, H., Marchio, A., Tiollais, P. & Dejean, A. *Nature* **330**, 667–670 (1987).
9. Hsu, T. Y. *et al. Cell* **55**, 627–635 (1988).
10. Fourrel, G. *et al. Nature* **347**, 294–298 (1990).
11. Wollersheim, M., Debelka, U. & Hofschneider, P. H. *Oncogene* **3**, 545–552 (1988).
12. Twu, J.-S. & Schloemer, R. H. *J. Virol.* **61**, 3448–3453 (1987).
13. Spandau, D. F. & Lee, C. H. *J. Virol.* **62**, 427–434 (1988).
14. Seto, E., Mitchell, P. J. & Yen, T. S. B. *Nature* **344**, 72–74 (1990).
15. Vogel, J., Hinrichs, S. H., Reynolds, R. K., Luciw, P. A. & Jay, G. *Nature* **335**, 606–611 (1988).
16. Nerenberg, M., Hinrichs, S. H., Reynolds, R. K., Khoury, G. & Jay, G. *Science* **237**, 1324–1329 (1987).
17. Green, J. E., Hinrichs, S. H., Vogel, J. & Jay, G. *Nature* **341**, 72–74 (1989).
18. Koike, K., Hinrichs, S. H., Issebacher, K. J. & Jay, G. *Proc. natn. Acad. Sci. U.S.A.* **86**, 5615–5619 (1989).
19. Small, J. A., Scangos, G., Cork, L., Jay, G. & Khoury, G. *Cell* **46**, 13–18 (1986).
20. Small, J. A., Khoury, G., Jay, G., Howley, P. & Scangos, G. A. *Proc. natn. Acad. Sci. U.S.A.* **83**, 8288–8292 (1986).
21. Fujiyama, A. *et al. Nucleic Acids Res.* **11**, 4601–4610 (1983).
22. Shaul, Y., Rutter, W. J. & Laub, O. *EMBO J.* **4**, 427–430 (1985).
23. Treinin, M. & Laub, O. *Molec. cell. Biol.* **7**, 545–548 (1987).
24. Kaneko, S. & Miller, R. H. *J. Virol.* **62**, 3979–3984 (1988).
25. Dejean, A., Lugassy, C., Zafrani, S., Tiollais, P. & Brechot, C. *J. gen. Virol.* **65**, 651–655 (1984).
26. Koike, K., Akatsuka, T. & Miyamura, T. *Virology* **163**, 233–235 (1988).
27. Klimke, F., Mayer, D. & Bannasch, P. *Carcinogenesis* **5**, 265–268 (1984).
28. Brooks, P. N. & Roe, F. J. C. in *Digestive System* (eds Jones, T. C., Mohr, U. & Hunt, R. D.) 47–52 (Springer, Berlin/Heidelberg, 1985).
29. Bannasch, P., Zerban, H. & Hacker, H. J. in *Digestive System* (eds Jones, T. C., Mohr, U. & Hunt, R. D.) 10–30 (Springer, Berlin/Heidelberg, 1985).
30. Bedkar, F. F. *Cancer Res.* **42**, 3918–3923 (1982).
31. Popp, J. A. in *Digestive System* (eds Jones, T. C., Mohr, U. & Hunt, R. D.) 39–46 (Springer, Berlin/Heidelberg, 1985).
32. Chisari, F. V. *et al. Science* **230**, 1157–1160 (1985).
33. Babinet, C., Farza, H., Morello, D., Hadchouel, M. & Pourcel, C. *Science* **230**, 1160–1163 (1985).
34. Burk, R. D., DeLoiz, J. A., Elawady, M. K. & Gerhart, J. D. *J. Virol.* **62**, 649–654 (1988).
35. Farza, H. *et al. J. Virol.* **62**, 4149–4152 (1988).
36. Araki, K. *et al. Proc. natn. Acad. Sci. U.S.A.* **86**, 207–211 (1989).
37. Lee, T.-H. *et al. J. Virol.* **64**, 5939–5947 (1990).
38. Cosman, D., Kress, M., Khoury, G. & Jay, G. *Proc. natn. Acad. Sci. U.S.A.* **79**, 4947–4951 (1982).
39. Kress, M., Cosman, D., Khoury, G. & Jay, G. *Cell* **34**, 189–196 (1983).

ACKNOWLEDGEMENTS. We thank F. V. Chisari for discussion and advice, C. Bieberich and J. Vogel for critical reading of the manuscript, and L. Ruiz and L. Rainone for help in its preparation. This work was supported by the NIH.

Regulation by phosphorylation of reversible association of a myristoylated protein kinase C substrate with the plasma membrane

Marcus Thelen, Antony Rosen, Angus C. Nairn & Alan Aderem*

The Rockefeller University, 1230 York Avenue, New York, New York, USA

PROTEIN kinase C (PKC) transduces receptor-mediated signals by phosphorylating membrane-bound substrates which then act as effectors of specific cellular responses¹. The myristoylated alanine-rich C kinase substrate (MARCKS) is a specific PKC substrate which has been implicated in macrophage activation, neurosecretion and growth factor-dependent mitogenesis^{2–5}. Myristoylation of MARCKS is required for effective binding to the plasma membrane⁶ where it colocalizes with PKC⁷. Here we report that PKC-dependent phosphorylation displaces MARCKS from the membrane and that its subsequent dephosphorylation is accompanied by its reassociation with the membrane. This cycle of phosphorylation-dependent membrane attachment and detachment of a myristoylated protein represents a novel mechanism of reversible membrane targeting. As MARCKS is a calmodulin- and actin-binding protein (ref. 8, and J. Hartwig *et al.*, manuscript submitted), the cycle of membrane attachment/detachment represents a mechanism through which PKC might reversibly regulate actin-membrane interaction.

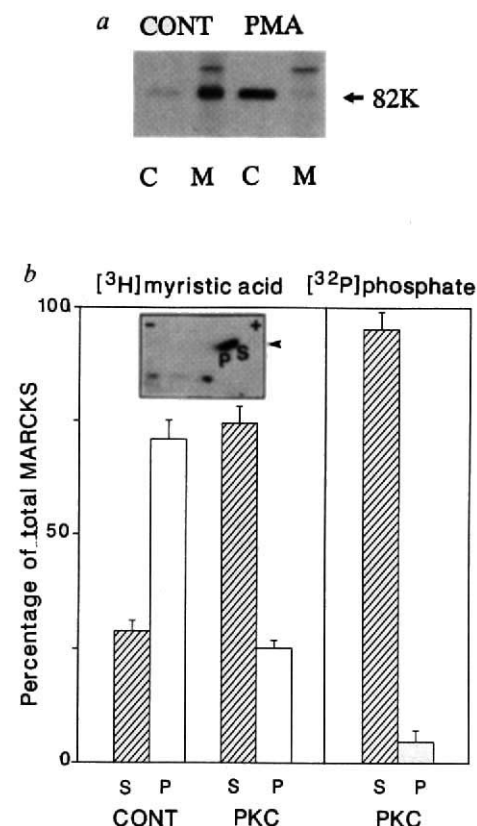
Treatment of neutrophils with phorbol myristate acetate (PMA) resulted in the rapid phosphorylation of MARCKS⁹ and in the concomitant displacement of all of the [³H]myristate-labelled protein from the membrane to the cytosol (Fig. 1a). Identical results were obtained when MARCKS was labelled biosynthetically with [³H]lysine or when total MARCKS was identified by immunoblotting (data not shown). Similarly, *in vitro* phosphorylation of isolated macrophage membranes with purified PKC resulted in the detachment of [³H]myristic acid-labelled MARCKS from the membrane, and this was associated with a shift to a more acidic pI (Fig. 1b). This acid shift of MARCKS is due to the incorporation of phosphate into the released protein (Fig. 1b, insert and right panel). MARCKS was not phosphorylated by the active catalytic subunit of cyclicAMP-dependent protein kinase or by calmodulin kinase I or II, and was not displaced from isolated membranes by these kinases (data not shown). These data show that the release of myristoylated MARCKS from the membrane is mediated by its phosphorylation by PKC and is not a consequence of its deacylation.

We next examined whether dephosphorylation of cytosolic MARCKS results in its reassociation with the membrane through its stably attached, myristic acid, membrane-targeting moiety. Reversibility of the membrane binding of MARCKS was investigated by activating neutrophils with the chemotactic agonist f-Met-Leu-Phe, a formyl peptide which stimulates transient PKC-dependent responses through a receptor-mediated, G protein-coupled pathway¹⁰. Treatment of neutrophils with f-Met-Leu-Phe resulted in the rapid, but transient, phosphorylation of MARCKS (Fig. 2). The protein was phosphorylated maximally 40 s after exposure of the cells to the agonist and phosphorylation had returned to basal levels after 4 min. All the phosphorylated protein was found in the cytosol (Fig. 2). As observed with PMA (Fig. 1a), f-Met-Leu-Phe-induced phosphorylation of MARCKS was accompanied by the displacement of myristic acid-labelled protein from the membrane to the cytosol (Fig. 2).

* To whom correspondence should be addressed.

FIG. 1 Activation of PKC results in the displacement of [3 H]myristate-labelled MARCKS from the plasma membrane. *a*, In control human neutrophils (CONT) 80–90% of labelled MARCKS (82K) is associated with the plasma membrane fraction (M). Stimulation with 100 nM PMA for 2 min results in the displacement of >90% of MARCKS from the membrane (M) to the cytosol (C). Note that the 90K (M, 90,000) acylated protein is not displaced from the membrane upon activation of PKC. *b*, PKC-induced displacement of MARCKS from isolated plasma membranes *in vitro*. When [3 H]myristate-labelled membranes from murine macrophages were incubated in the absence of PKC (CONT), 71% of myristoylated MARCKS remained with the membrane fraction (P), whereas 29% was found in the supernatant (S). Addition of active PKC resulted in the release of MARCKS from the membrane fraction, such that 75% of the myristoylated protein was observed in the supernatant (S), and only 25% was found in the pellet (P) (left panel). MARCKS shifts towards a more acidic pI upon its phosphorylation and release from the membrane (the inset, left panel, shows a mixture of MARCKS obtained from both membrane (less acidic) (P) and supernatant (more acidic) (S) fractions). In experiments in which [γ - 32 P]ATP was used, 95% of phosphorylated MARCKS was found in the supernatant (right panel). The data represents the mean $\pm$ s.d. of six experiments.

METHODS. In *a*, neutrophils were treated with 100 U ml $^{-1}$ recombinant human TNF α and labelled with [3 H]myristic acid as previously described⁹. Following stimulation, the cells were resuspended in relaxation buffer without ATP¹⁶, disrupted by nitrogen cavitation¹⁶, and centrifuged on a discontinuous sucrose gradient⁷. Fractions containing plasma membrane or cytosol were analysed by SDS-PAGE and myristoylated proteins were visualized by fluorography⁹. The identity of MARCKS was confirmed by immunoprecipitation⁹. In *b*, plasma membranes from lipopolysaccharide (100 ng ml $^{-1}$) primed, [3 H]myristate-labelled² murine macrophages were prepared⁷ and treated for 2 min at 30 °C with buffer (CONT) or purified PKC (PKC)¹⁷. The incubation was terminated by centrifugation (15 min, 4 °C, 436,000g) and MARCKS was immunoprecipitated from the pellet (P) and supernatant (S) as described previously². The relative amount of [3 H]myristate-labelled MARCKS in either fraction was determined from fluorographs by laser densitometry. The pI of [3 H]myristoylated, nonphosphorylated MARCKS (from CONT, P) and of [3 H]myristoylated, phosphorylated MARCKS (from PKC, S) was determined by 2D-IEF/SDS-PAGE as described⁹. In parallel experiments, [3 H]myristic



acid was omitted and the isolated membranes were phosphorylated using [γ - 32 P]-ATP¹⁷ (right panel).

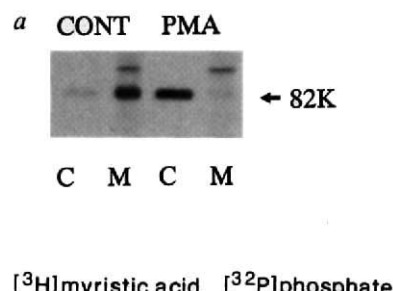
Thus the level of membrane-bound MARCKS decreased from 75% to 35% after 40 s stimulation with f-Met-Leu-Phe, and the amount of cytosolic protein increased from 25% to 65% over the same time course (Fig. 2). After 40 s stimulation with f-Met-Leu-Phe, the equilibrium between kinase and phosphatase shifted to favour the dephosphorylation of MARCKS (Fig. 2). As dephosphorylation proceeded there was a concomitant reassociation of the protein with the membrane, such that 65% of MARCKS was membrane-bound after 4 min when its phosphorylation had decreased to basal levels (Fig. 2). The cycle of membrane release and attachment was not influenced by cycloheximide, indicating that *de novo* synthesis of MARCKS did not account for the increase in the membrane-bound form of the protein observed upon its dephosphorylation (data not shown).

FIG. 2 Transient, agonist-dependent phosphorylation of MARCKS is associated with a cycle of detachment from and attachment to the plasma membrane. Treatment of human neutrophils with 500 nM f-Met-Leu-Phe resulted in maximal levels of phosphorylated MARCKS within 40 s, after which phosphorylation decreased (lower panel). More than 90% of phosphorylated MARCKS was found in the cytosol (C). In unstimulated cells about 70–80% of [3 H]myristoylated MARCKS associated with the membrane (M) whereas 20–30% are found in the cytosol (C). f-Met-Leu-Phe-induced phosphorylation of MARCKS was associated with a translocation of the protein to the cytosol such that 65% of [3 H]myristic acid-labelled protein was cytosolic and 35% remained membrane bound. As dephosphorylation proceeded [3 H]myristate-labelled MARCKS reassociated with the membrane such that 65% of the labelled protein was again membrane-bound 240 s after addition of the stimulus. The level of cytosolic MARCKS in unstimulated cells varied between 10–25% in different preparations, but higher levels of cytosolic MARCKS were always accompanied by increased basal levels of phosphorylation.

METHODS. Human neutrophils were labelled with [3 H]myristic acid or with

Okadaic acid, a specific phosphatase inhibitor^{11,12}, prevented the dephosphorylation of MARCKS in neutrophils (shown at the 3-min time point; Fig. 3, left panel). Concomitantly, okadaic acid also inhibited the reassociation of [3 H]myristic acid-labelled MARCKS with the membrane, further strengthening the contention that dephosphorylation of the protein is required for effective membrane binding (Fig. 3, right panel). As okadaic acid shifts the equilibrium between the kinase and the phosphatase in f-Met-Leu-Phe-stimulated neutrophils, the observation that the level of phosphorylated and of [3 H]myristic acid-labelled MARCKS in the cytosol was slightly enhanced in okadaic acid-treated cells was anticipated.

As the binding of f-Met-Leu-Phe to its receptor transiently stimulates the phosphorylation of MARCKS (Fig. 2) we investigated whether a specific antagonist, boc-Met-Leu-Phe, shifted



[32 P]phosphate⁹ and were then fractionated as described in Fig. 1. In the lower panel MARCKS was immunoprecipitated from the Nonidet P-40 (NP-40)-solubilized plasma membrane fraction and from the cytosol, and visualized by autoradiography⁹.

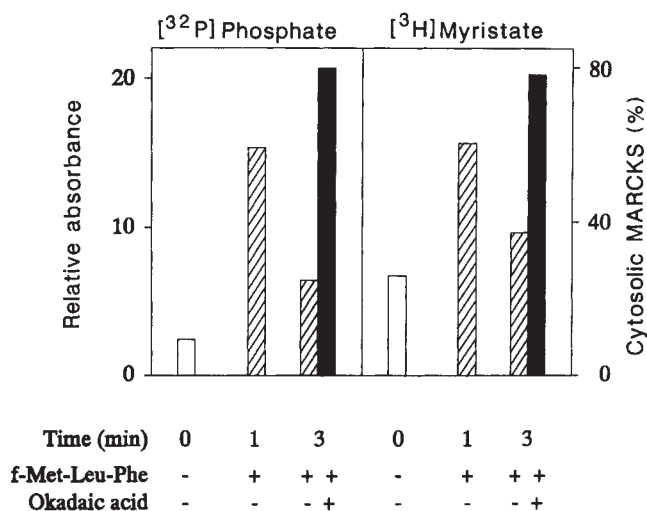


FIG. 3 Okadaic acid prevents the reassociation of MARCKS with the plasma membrane of human neutrophils. Stimulation of neutrophils with 500 nM f-Met-Leu-Phe (hatched bars) induced the phosphorylation (left panel) and membrane displacement (right panel) of MARCKS within 1 min. When the cells were preincubated with 1 μ M okadaic acid (filled bars) dephosphorylation was inhibited (left panel) and the [³H]myristoylated protein did not reassociate with the membrane (right panel). The data are representative of six experiments which did not vary by more than 15%. METHODS. Neutrophils were labelled with [³²P]phosphate or with [³H]myristic acid (see Fig. 2) and were treated with 1 μ M okadaic acid or its carrier DMSO (0.8% f.c.) for 5 min before stimulation with 500 nM f-Met-Leu-Phe. After fractionation MARCKS was immunoprecipitated as described above. All the [³²P]labelled MARCKS was found in the cytosol and the resulting densitometric absorbance values derived by scanning the autoradiograph is plotted in relative units (left panel). The [³H]myristate-labelled protein in the membrane and cytosolic fractions was measured by densitometry of the fluorographs and is expressed as the percentage of cytosolic MARCKS (right panel).

the equilibrium of f-Met-Leu-Phe-induced phosphorylation of MARCKS and consequently influenced its association with the membrane. Addition of excess boc-Met-Leu-Phe to neutrophils 1 min after stimulation with f-Met-Leu-Phe resulted in a rapid decrease in the level of phosphorylated MARCKS (70% decrease within 1 min) (Fig. 4, upper panel), which was paralleled by a concomitant increase in the rate of reassociation of the protein with the membrane (Fig. 4, lower panel). In addition, the data suggest a tight coupling between receptor occupancy and the activation of PKC, as evidenced by the rapid reversal in the level of phosphorylation of the specific PKC substrate (MARCKS) after displacement of the agonist.

The capacity of MARCKS to shuttle reversibly between the membrane and cytosol in a phosphorylation-dependent manner, together with our observation that MARCKS is located in focal contacts at the substrate-adherent plasma membrane⁷, suggest that the protein associates with the membrane through a receptor, rather than through nonspecific insertion of the myristic acid in the lipid bilayer. Evidence for a myristoyl-protein receptor for the proto-oncogene products p60^{src} and p56^{lck} has been presented¹³⁻¹⁵, and as these two proteins are also PKC substrates, it would be interesting to determine whether their capacity to bind to the membrane is similarly modified by phosphorylation.

MARCKS binds both actin (J. Hartwig *et al.*, manuscript submitted) and calmodulin⁸ and colocalizes with vinculin, talin and PKC at the substrate-adherent surface of macrophage filopodia⁷. Treatment of macrophages with phorbol esters results in the displacement of MARCKS from these focal contacts and in rearrangement of the actin cytoskeleton⁷. The cycles of membrane attachment and detachment of MARCKS might therefore provide a PKC-sensitive, reversible crossbridge between the actin cytoskeleton and the plasma membrane. Such a reversible association of actin with the substrate-adherent membrane is a

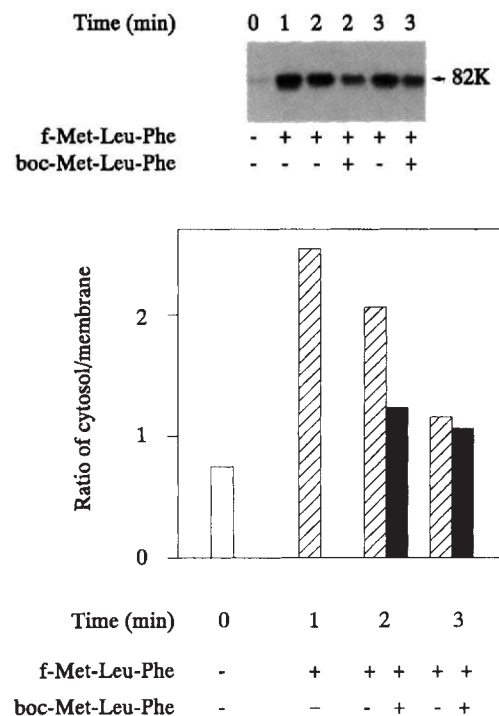


FIG. 4 A receptor antagonist accelerates the dephosphorylation of MARCKS and its reassociation with the membrane. Upper panel, transient phosphorylation of MARCKS was stimulated with 10 nM f-Met-Leu-Phe. Addition of the antagonist boc-Met-Leu-Phe (10 μ M) one minute after the agonist markedly accelerated the dephosphorylation of MARCKS. Lower panel, f-Met-Leu-Phe-dependent phosphorylation of MARCKS was accompanied by the displacement of the [³H]myristate-labelled protein from the membrane to the cytosol. Accelerated dephosphorylation induced by boc-Met-Leu-Phe (see upper panel) was mirrored by the more rapid reassociation of [³H]myristate-labelled MARCKS with the membrane (lower panel). Note that the lower concentration of the agonist (10 nM f-Met-Leu-Phe) used in this experiment resulted in a more transient phosphorylation and membrane displacement of MARCKS than those observed with the higher doses used in Figs 2 and 3. The data shown are representative of five similar experiments. METHODS. Neutrophils were labelled with [³²P]phosphate and [³H]myristic acid as described for Fig. 2 and were then stimulated with f-Met-Leu-Phe for the times indicated. The antagonist boc-Met-Leu-Phe was added 1 min after the stimulation. [³²P]phosphorylated MARCKS was immunoprecipitated⁹ as described in Fig. 3. [³H]myristate-labelled neutrophils were fractionated as described in the legend to Fig. 2.

prerequisite for directed cellular locomotion during chemotaxis, and it is significant that the bacterially derived agent f-Met-Leu-Phe induces chemotaxis in neutrophils and causes MARCKS to shuttle between the membrane and the cytosol. □

Received 14 December 1990; accepted 29 March 1991.

- Nishizuka, Y. *Nature* **334**, 661-665 (1988).
- Aderem, A. A. *et al.* *Nature* **332**, 362-364 (1988).
- Wu, W. S., Walaas, S. I., Nairn, A. C. & Greengard, P. *Proc. natn. Acad. Sci. U.S.A.* **79**, 5249-5253 (1982).
- Rozengurt, E., Rodriguez-Pena, M. & Smith, K. A. *Proc. natn. Acad. Sci. U.S.A.* **80**, 7244-7248 (1983).
- Blackshear, P. J., Wen, L., Glynn, B. P. & Witters, L. A. *J. biol. Chem.* **261**, 1459-1469 (1986).
- Graff, J. M., Gordon, J. I. & Blackshear, P. J. *Science* **246**, 503-506 (1989).
- Rosen, A., Keenan, K. F., Thelen, M., Nairn, A. C. & Aderem, A. A. *J. exp. Med.* **172**, 1211-1215 (1990).
- Graff, J. M., Young, T. M., Johnson, J. D. & Blackshear, P. J. *J. biol. Chem.* **264**, 21818-21823 (1989).
- Thelen, M., Rosen, A., Nairn, A. C. & Aderem, A. A. *Proc. natn. Acad. Sci. U.S.A.* **87**, 5603-5607 (1990).
- Dewald, B., Thelen, M. & Baggiolini, M. *J. biol. Chem.* **263**, 16179-16184 (1988).
- Blaloe, J. C. & Takai, A. *Biochem. J.* **256**, 283-290 (1988).
- Haystead, T. A. J. *et al.* *Nature* **337**, 78-81 (1989).
- Resh, M. D. *Cell* **58**, 281-286 (1989).
- Resh, M. D. & Ling, H. *Nature* **346**, 84-86 (1990).
- Veilleux, A., Bookman, M. A., Horak, E. M. & Bolen, J. B. *Cell* **55**, 301-308 (1988).
- Borregaard, N., Heiple, J. M., Simons, E. R. & Clark, R. A. *J. Cell Biol.* **97**, 52-61 (1983).
- Rosen, A., Nairn, A. C., Greengard, P., Cohn, Z. A. & Aderem, A. A. *J. biol. Chem.* **264**, 9118-9121 (1989).

ACKNOWLEDGEMENTS. M.T. is a fellow of the Deutsche Forschungsgemeinschaft. A.R. is a recipient of a Cancer Research Institute/Richard C. and Susan B. Ernst Fellowship, and A.A. is an established investigator of the American Heart Association. This work was supported by the NIH.

Restored expression of major histocompatibility class I molecules by gene transfer of a putative peptide transporter

Thomas Spies* & Robert DeMars†

* Division of Tumor Virology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA

† Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706, USA

CYTOTOXIC T lymphocytes recognize antigen-derived peptides bound to major histocompatibility complex (MHC) class I molecules with which they assemble in the endoplasmic reticulum or in an undefined subcompartment¹⁻¹⁰. There is genetic evidence that the peptides that are products of cytosolic protein degradation are transported into this compartment by a peptide supply factor (PSF), encoded in the MHC class II region¹¹. Like the corresponding genes *RING4*, *HAM1* and *mtp1* (refs 12-14), *PSF* is related to the multidrug-resistance family of transporters¹¹ and may be a peptide pump, as translocation of peptides across membranes must occur independently of the secretory pathway¹⁵. There is, however, no functional evidence for this role so far. Here we report gene transfer experiments showing that expression of PSF complementary DNA in the human lymphoblastoid cell line mutant 721.134 (refs 11, 16, 17) restores normal levels of surface HLA-A2 and -B5. No similar effect was observed in 721.174 mutant cells, in which a homozygous deletion includes *PSF* among several other closely linked genes¹¹. At least one of these genes may therefore also be required for PSF function.

Expression of PSF cDNA in LCL mutant 721.134 cells was pursued after failure to restore PSF function by cosmid transfection. The selected PSF cDNA (Y3-1) is 2,667 base pairs (bp) long, includes 72 bp upstream of a putative translation initiation codon and terminates 2 bp before the polyadenylation site (our unpublished data). The translation product is identical to the amino-acid sequence encoded by *RING4* (ref. 12). The PSF cDNA was inserted into pcDNA I/NEO (Invitrogen), under transcriptional control of the cytomegalovirus promoter and enhancer. Transfer of pcDNA I/NEO-PSF into .134 cells and selection for the neomycin (G418) resistance marker yielded 54 independent transfectant cell populations. These were examined for surface expression of HLA-B5 by flow cytometry. In 27 isolates, 3-30% of cells bound the 4D12 monoclonal antibody, which recognizes several class I alleles including HLA-B5 but not HLA-A2^{18,19} (Fig. 1A, c-f). The amount of HLA-B5 on these cells was equivalent to the wild-type expression level on LCL 721.45 cells, from which mutant .134 originated (Fig. 1A, a). Identical antibody-binding profiles were obtained with monoclonal antibody B1.23.2, which is pan-reactive anti-HLA-B²⁰ (data not shown). The presence of large fractions of HLA-B5⁺ cells was related to the high frequency of gene transfer; each single-well isolate contained multiple transfectant clones, in most of which the PSF cDNA may have been deleted or rearranged upon chromosomal integration. However, individual HLA-B5⁺ cell clones obtained by limiting dilution from two transfectant cell populations, 1-C4 and 1-A4 (Fig. 1A, e, f), showed stable expression of HLA-B5 after at least two months of culture in the absence of drug selection (Fig. 2a-c). In contrast, .134 cells and .134 cell populations transfected with pcDNA I/NEO invariably lacked surface HLA-B5 (Fig. 1A, b).

In a parallel experiment, the PSF cDNA was ligated into RSV.5(gpt)²¹, in which cDNA transcription is driven by the Rous sarcoma virus 5' long terminal repeat. In RSV.5(gpt), the

guanine phosphoribosyltransferase gene confers resistance to mycophenolic acid. After transfection of .134 cells with RSV.5(gpt)-PSF and selection, 42 distinct transfectant cell populations were isolated, of which 17 had proportions of 2-95% of cells positive for surface HLA-B5 (Fig. 1B, c-f). All of these displayed at least as much HLA-B5 as the control .45 cells (Fig. 1B, a). In one isolate (2-W), the level of expression was significantly higher (Fig. 1B, f).

Although HLA-B5 is absent from the surface of .134 cells, HLA-A2 is present at about 40% of the parental .45 level of expression (Fig. 1C, a, b)¹⁷. Concomitant with the HLA-B5⁺ phenotype, all the RSV.5(gpt)-PSF transfectant cell populations

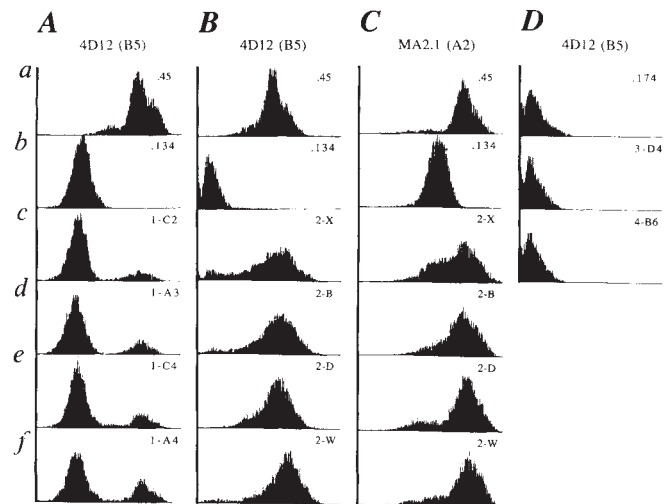


FIG. 1 PSF cDNA transfer restores cell surface expression of HLA-B5 and -A2 in LCL mutant .134, but not in LCL mutant .174. A, Flow cytometry of .134 cell populations transfected with pcDNA I/NEO-PSF. In c-f, profiles show binding of monoclonal antibody 4D12 (anti-HLA-B5)¹⁸ to 15-30% of cells selected for resistance to neomycin (G418). Fluorescence intensity was equal to the positive control, LCL .45 cells (a). Little if any HLA-B5 was detected on the surface of .134 cells (b). B, In populations of .134 cells transfected with RSV.5(gpt)-PSF and selected for resistance to mycophenolic acid, 80-95% of cells bound monoclonal antibody 4D12 (c-f). In f, fluorescence intensity was significantly higher than on .45 cells (a). C, Isolates shown in B were stained with monoclonal antibody MA2.1 (anti-HLA-A2)²². D, Absence of binding of 4D12 to LCL .174 cells transfected with pcDNA I/NEO-PSF (b) or RSV.5(gpt)-PSF (c) and selected for drug-resistance and expression of PSF mRNA (Fig. 2a). Profiles b and c are representative of 60 independent polyclonal isolates containing altogether a minimum of 180 independent transfectants.

METHODS. PSF cDNA Y3-1 was inserted into pcDNA I/NEO (Invitrogen) and RSV.5(gpt)²¹ using flanking polylinker restriction sites *HindIII* and *NotI* and *XbaI*, respectively. Transfection of the mutant LCLs was carried out with an electroporation apparatus (constructed by D. Yee, Medical Electronics Laboratory, University of Wisconsin), using 1250 V, three capacitor banks, minimum rise time and maximum fall time. Cells were cultured at about $5 \times 10^5 \text{ ml}^{-1}$ for at least 2 days before electroporation. Cells (5×10^6) were suspended in 0.5 ml of culture medium (RPMI 1640/15% FCS) in electroporation cuvettes placed in an ice water bath. After electroporation, cells were maintained at room temperature for 10 min, resuspended in fresh culture medium and distributed into multi-well plates. The amount of DNA per cuvette was 20 μg and 3 μg in the experiments A and B-D, respectively. Selection started on day 5 after electroporation, with G418 (Gibco) ($400 \mu\text{g ml}^{-1}$) for pcDNA I/NEO-PSF and with mycophenolic acid (Sigma) ($10 \mu\text{g ml}^{-1}$) in medium containing xanthine ($10 \mu\text{g ml}^{-1}$)²⁴ for RSV.5(gpt)-PSF. After 2 weeks of selection, drug-resistant cell populations proliferated in many wells and were expanded individually. Analysis of transfectant cell populations by flow cytometry began 5 weeks after electroporation. Cells were maintained at $3-8 \times 10^5 \text{ ml}^{-1}$ for 2-3 days and stained by indirect immunofluorescence as described¹⁷. Following binding of antibodies 4D12 and MA2.1, FITC-conjugated goat anti-mouse immunoglobulin (Coulter) was used as the second-step reagent. Selected cell populations were sampled ($3-5 \times 10^3$ cells) and analysed two or three times at intervals of about 1 week with concordant results using a Coulter Electronics Epics C flow cytometer.

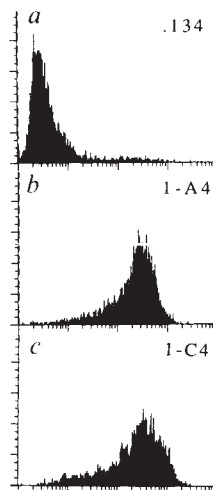


FIG. 2 Stable surface expression of HLA-B5 in cell clones of LCL mutant .134 transfected with pcDNA I/NEO-PSF. Samples shown are representative of 12 isolates obtained by limiting dilution cloning from the 1-C4 and 1-A4 transfectant cell populations (Fig. 1*A*, *e* and *f*). Cells were grown in the absence of G418 for at least two months. Cells were stained with monoclonal antibody 4D12 and FITC-goat anti-mouse immunoglobulin and examined with a Becton Dickinson FACScan flow cytometer.

showed increased binding of monoclonal antibody MA2.1, which is monospecific anti-HLA-A2²² (Fig. 1*C*, *a-f*). Surface staining was at least as intense as in .45 cells (Fig. 1*C*, *a*). Concordant data were obtained from pcDNA I/NEO-PSF transfectant HLA-B5⁺ isolates and when binding of monoclonal antibody BB7.2²³ was used to measure surface HLA-A2 (data not shown). Thus, quantitatively normal expression of HLA-A2 and -B5 was restored in mutant .134 cells. In addition, the results suggest accurate folding at least of HLA-A2 chains, as recognition by allele-specific monoclonal antibodies may be conformation-sensitive.

In all the HLA-B5⁺ .134 transfectant cell populations studied, a major PSF RNA transcript was longer than the PSF messenger RNA of 2.7 kilobases (kb) in control .45 cells, because of fusion with vector-derived polyadenylation sequences (Fig. 3*a*). A minor hybridizing RNA of 1.6 kb in pcDNA I/NEO-PSF transfectant cells was probably an artefact caused by DNA rearrangement or abortive RNA splicing. No RNA of 2.7 kb was detected in any transfectant isolate (Fig. 3*a*). This reinforces the evidence that surface expression of HLA-A2 and -B5 in .134 cells was restored as a result of PSF cDNA transfer and that a mutation abolishing transcription of the *PSF* gene is

responsible for the mutant .134 phenotype¹¹.

In several of the .134 transfectant cell populations, the relative amounts of PSF mRNA were independent of the levels of HLA-A2 and -B5 surface expression (Fig. 3). This may be related to heterogeneity in the polyclonal cell isolates containing functional and non-functional PSF mRNA and to positional effects of chromosomal insertion sites. In addition, differential cell proliferation rates may have resulted in shifts in the clonal composition of some cell populations, for example 1-C2 (Fig. 3*a*).

In LCL mutant 721.174, defective surface expression of HLA-A2 and -B5 is associated with a 1-megabase homozygous genomic deletion including *PSF*¹¹. Transfection of .174 cells with pcDNA I/NEO-PSF and RSV.5(gpt)-PSF yielded a total of 60 transfectant cell populations, all of which failed to bind monoclonal antibodies 4D12 and B1.23.2 (Fig. 1*D*, *a-c*). This is statistically significant, as HLA-B5⁺ cells occurred in about 50% of the .134 transfectant cell populations. In addition, vector-derived PSF mRNA was present in 16 of 36 .174 transfectant isolates (Fig. 3*a*). Thus, PSF gene transfer alone is not sufficient to restore surface expression of class I molecules in mutant .174 cells. At least one additional gene among several candidates identified previously near *PSF* may be required¹¹. This apparent polygenicity may suggest that distinct subunits cooperate in a functional peptide transporter. □

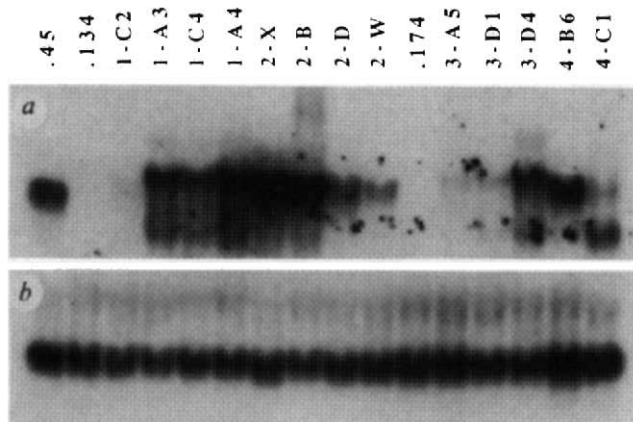


FIG. 3 Vector-derived PSF mRNA in .134 and .174 transfectant cell populations. *a*, Blot hybridization of total RNA preparations (20 µg per lane) from isolates shown in Fig. 1 with a PSF cDNA probe. Transcripts from pcDNA I/NEO-PSF and RSV.5(gpt)-PSF are 840 bp and 120 bp longer than the 2.7 kilobase (kb) PSF mRNA in control .45 cells, respectively, due to addition of polylinker and polyadenylation sequences. Like sample 3-D4, 3-A5 and 3-D1 are .174 isolates transfected with pcDNA I/NEO-PSF. In sample 4-C1, RSV.5(gpt)-PSF is rearranged in .174 cells. *b*, Parallel blot hybridization with a probe for HLA-B25 to control for loading. Autoradiography was for 4 days in *a* and 16 h in *b*. Methods as previously described²⁵. See text for further explanation.

Received 4 March; accepted 3 April 1991.

1. Townsend, A. R. M., Gotch, F. M. & Davey, J. *Cell* **42**, 457-467 (1985).
2. Morrison, L. A., Lukacher, A. E., Braciale, V. L., Fan, D. P. & Braciale, T. J. *J. exp. Med.* **163**, 903-921 (1986).
3. Bjorkman, P. J. *et al.* *Nature* **329**, 512-516 (1987).
4. Moore, M. W., Carbone, F. R. & Bevan, M. J. *Cell* **54**, 777-785 (1988).
5. Van Bleek, G. M. & Nathenson, S. G. *Nature* **348**, 213-216 (1990).
6. Townsend, A. *et al.* *Nature* **340**, 443-448 (1989).
7. Cerundolo, V. *et al.* *Nature* **345**, 449-452 (1990).
8. Salter, R. D. & Cresswell, P. *EMBO J.* **5**, 943-949 (1986).
9. Townsend, A. *et al.* *Cell* **62**, 285-295 (1990).
10. Schumacher, T. N. M. *et al.* *Cell* **62**, 563-567 (1990).
11. Spies, T. *et al.* *Nature* **348**, 744-747 (1990).
12. Trowsdale, J. *et al.* *Nature* **348**, 741-744 (1990).
13. Monaco, J. J., Cho, S. & Attaya, M. *Science* **250**, 1723-1726 (1990).
14. Deverson, E. V. *et al.* *Nature* **348**, 738-741 (1990).
15. Townsend, A. R. M., Bastin, J., Gould, K. & Brownlee, G. G. *Nature* **324**, 575-577 (1986).
16. DeMars, R., Chang, C. C., Shaw, S., Reitnauer, P. J. & Sondel, P. M. *Hum. Immun.* **11**, 77-79 (1984).
17. DeMars, R. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **82**, 8183-8187 (1985).
18. Haynes, B. F., Reisner, E. G., Hemler, M. E., Strominger, J. L. & Eisenbarth, G. S. *Hum. Immun.* **4**, 273-285 (1982).
19. Mann, D. L. *et al.* *Nature* **305**, 58-60 (1983).
20. Shimizu, Y. *et al.* *Molec. cell Biol.* **6**, 1074-1087 (1986).
21. Long, E. O. *et al.* *Hum. Immun.* (in the press).
22. Ways, J. P., Rothbard, J. B. & Parham, P. *J. Immun.* **137**, 217-222 (1986).
23. Brodsky, F. M., Parham, P., Barnstable, C. J., Crumpton, M. J. & Bodmer, W. F. *Immun. Rev.* **47**, 3-61 (1979).
24. Mulligan, R. C. & Berg, P. *Proc. natn. Acad. Sci. U.S.A.* **78**, 2072-2076 (1981).
25. Spies, T., Blanck, G., Bresnahan, M., Sands, J. & Strominger, J. L. *Science* **243**, 214-217 (1989).

ACKNOWLEDGEMENTS. We thank M. Hemier, E. Long, B. Malissen and S. Ono for gifts of plasmid vectors and monoclonal antibodies and J. Petersen for making the RSV.5(gpt)-PSF subclone. This research was supported by the German Rheumatism Research Center (Berlin) and the NIH.

Characterization of a murine gene expressed from the inactive X chromosome

Giuseppe Borsani, Rossana Tonlorenzi*,
M. Christine Simmler†, Luisa Dandolo†,
Danielle Arnaud†, Valeria Capra*, Markus Grompe,
Antonio Pizzuti, Donna Muzny, Charles Lawrence‡,
Huntington F. Willard§, Philip Avner†
& Andrea Ballabio||

Institute for Molecular Genetics and ‡ Department of Cell Biology,
Baylor College of Medicine, Houston, Texas 77030, USA

† Unité de Génétique Moléculaire Marine, Institut Pasteur,
25 rue du Docteur Roux, F-75724 Paris Cedex 15, France

§ Department of Genetics, Stanford University School of Medicine,
Stanford, California 94305, USA

IN mammals, equal dosage of gene products encoded by the X chromosome in male and female cells is achieved by X inactivation. Although X-chromosome inactivation represents the most extensive example known of long range *cis* gene regulation, the mechanism by which thousands of genes on only one of a pair of identical chromosomes are turned off is poorly understood. We have recently identified a human gene (*XIST*) exclusively expressed from the inactive X chromosome¹. Here we report the isolation and characterization of its murine homologue (*Xist*) which localizes to the mouse X inactivation centre region and is the first murine gene found to be expressed from the inactive X chromosome. Nucleotide sequence analysis indicates that *Xist* may be associated with a protein product. The similar map positions and expression patterns for *Xist* in mouse and man suggest that this gene may have a role in X inactivation.

There is genetic evidence that a single locus, the X inactivation centre (*XIC*) in man and the X controlling element (*Xce*) in mouse, is required in *cis* for inactivation to occur²⁻⁵. The chromosomal location of this locus in band Xq13 in man and bands D/E in the mouse X chromosome has recently been refined⁶⁻⁹. When the human *XIST* complementary DNA clone 14A (ref. 1) was hybridized under conditions of reduced stringency to DNAs of mice from an extensively typed interspecific backcross panel^{10,11}, restriction fragment length polymorphisms allowed localization of the corresponding locus to the central region of the mouse X chromosome in which *Xce* resides. As shown in Fig. 1a and c, the recombination breakpoints of the X chromosome in backcross animals 74, 114 and 194 establish the murine *Xist* locus as distal to *Ccg-1/Phka*, and those of animals 61 and 172 establish it as lying proximal to *Pgk-1*, consistent with the order of homologous loci in the human (*CCGL-XIST-PGKI*)⁶. Genetic analysis gives a recombination distance of 1.2+/-0.8 centimorgans between *Xist* and the *Ccg-1/Phka* loci (250 animals tested).

To isolate the murine *Xist* homologue, we hybridized the human *XIST* cDNA clone 14A to a cDNA library in λ ZAPII from thymus RNA of female mice (Stratagene). The longest of the isolated cDNA clones, MR20 (3.1 kilobases (kb)), gave results on the backcross panel indistinguishable from those found for the X-linked sequences detected with the human probe 14A (Fig. 1b and c) and consequently mapped it between the *Ccg-1/Phka* and *Pgk-1* loci. Independent confirmation of the mapping of the *Xist* locus to this region of the mouse X chromosome has been obtained from a series of somatic cell hybrids carrying various deletions of the mouse X chromosome¹¹ and from the use of a large panel of irradiation fusion somatic cell

hybrids obtained after 50 krad irradiation¹² (data not shown). Combined analysis of the deletion and meiotic mapping data establishes the order: centromere-*Ccg-1-Phka-Xist-DXPas19-Pgk-1*-telomere.

Expression of the *Xist* gene has been studied by polymerase chain reaction amplification of reverse-transcribed RNA

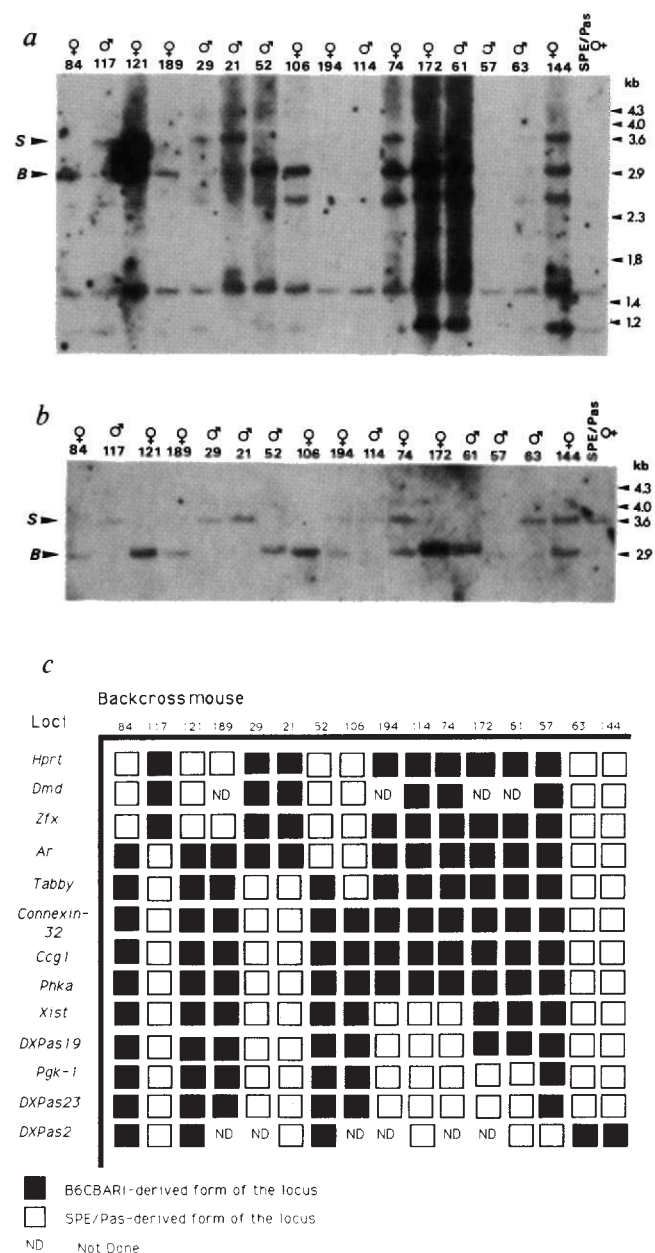


FIG. 1 Meiotic mapping of the *Xist* gene. **a**, DNAs from animals of an interspecific backcross, involving B6CBARI and inbred mouse strain SPE/Pas derived from *Mus spretus*, were hybridized to the human *XIST* cDNA 14A and the mouse *Xist* cDNA MR20. Under the conditions of reduced hybridization stringency used in the experiments with the human probe 14A, additional weak bands that do not apparently localize to the X chromosome are also observed. **b**, Hybridization pattern of the same panel with the mouse cDNA probe MR20. **c**, Pedigree analysis of individual backcross progeny from the *Mus spretus* backcross showing the position of the *Xist* locus. S, SPE/Pas-derived alleles; B, B6CBARI-derived alleles.

METHODS. For Southern blotting, the human probe 14A was hybridized in sodium phosphate-SDS buffer at 65 °C according to the method of Church and Gilbert¹⁷, followed by washing with 5 × SSC at 50 °C. Mouse probe MR20 was hybridized under identical conditions but washed with 2 × SSC at 68 °C. All probes were labelled by random priming¹⁸ and blots were exposed after hybridization in the presence of intensifying screens for various periods.

* On leave from G. Gaslini Institute, Genoa, Italy.
|| To whom correspondence should be addressed.

(RT-PCR). Figure 2a shows the RT-PCR amplification of liver RNA samples from three male and three female mice, using oligonucleotide primers derived from the mouse *Xist* cDNA sequence. Only samples corresponding to female RNAs gave any product, indicating that *Xist*, like its human homologue, is exclusively expressed from the inactive X chromosome. Levels of *Xist* RNA in XO and XY mouse liver were ~1,000-fold lower than those detected in the XX female samples containing an inactive X chromosome (Fig. 2b). The decreased transcription in our XO female samples, in which the X chromosome is of paternal origin, rules out the possibility that *Xist* transcriptional activity is controlled by paternal/maternal imprinting. Northern blot analysis of RNAs from liver samples of male and female mice gave a continuous smear only in the lanes corresponding to female samples (data not shown), as observed for the human *XIST* gene¹.

Figure 3a shows the PCR amplification of reverse-transcribed RNA, genomic DNA, cosmid clone MB4-14 (spanning the entire 3.1 kb of sequence contained in the cDNA clone MR20) and of clone MR20. The results demonstrate complete colinearity between the cDNA and the genomic DNA sequence, indicating that clone MR20 is a 3.1-kb unspliced transcript. Twenty-two additional *Xist* cDNA clones were identified using MR20 as a probe and characterized by restriction mapping, PCR amplification and hybridization of oligonucleotides designed from the sequence of clone MR20. The hybridization of nine oligonucleotides regularly spaced throughout the sequence of clone MR20 indicates the absence of alternative splicing, a conclusion based on the comparison of the length of each clone

with its hybridization pattern (Fig. 3b). These findings differ from those for the human gene, in which an extremely high frequency of alternative splicing is observed in the equivalent region¹, and establish a clear difference in the splicing patterns of the human and mouse *Xist* genes.

Figure 4a shows the nucleotide sequence of 4.1 kb of the *Xist* gene (deposited in the GenBank/EMBL Data Library, accession number X59289). Sequence analysis revealed an open reading frame 894 base pairs (bp) long located at the 5' end of the portion of the *Xist* gene characterized so far. The corresponding amino-acid sequence is rich in leucine and cysteine (25% and 12%, respectively). The putative protein is extremely hydrophobic, so possibly represents a membrane-bound domain. Analysis of the sequence revealed the presence of a 24-bp motif repeated eight times in the first 270 bp of the DNA sequence, reflected in a repeat of eight amino acids in the predicted protein (Fig. 4a). Screening of the GenBank sequence database using the *Xist* nucleotide sequence and of the Swiss-protein, PIR and GenPept databases using the predicted amino-acid sequence did not show significant similarity to any other sequence. We have also looked specifically for the repeat motif in the databases and have found no similar sequence. Although the long open reading frame strongly suggests that an *Xist* protein product exists, further studies are needed to provide a definitive answer on the protein-encoding capability of the *Xist* gene. The murine *Xist* sequence was compared with the longest unspliced stretch of cDNA sequence from the human *XIST* gene (J. Rupert, C. Brown and H. F. W., unpublished data). A homologous region of ~500 bp was found in a region corresponding to bases

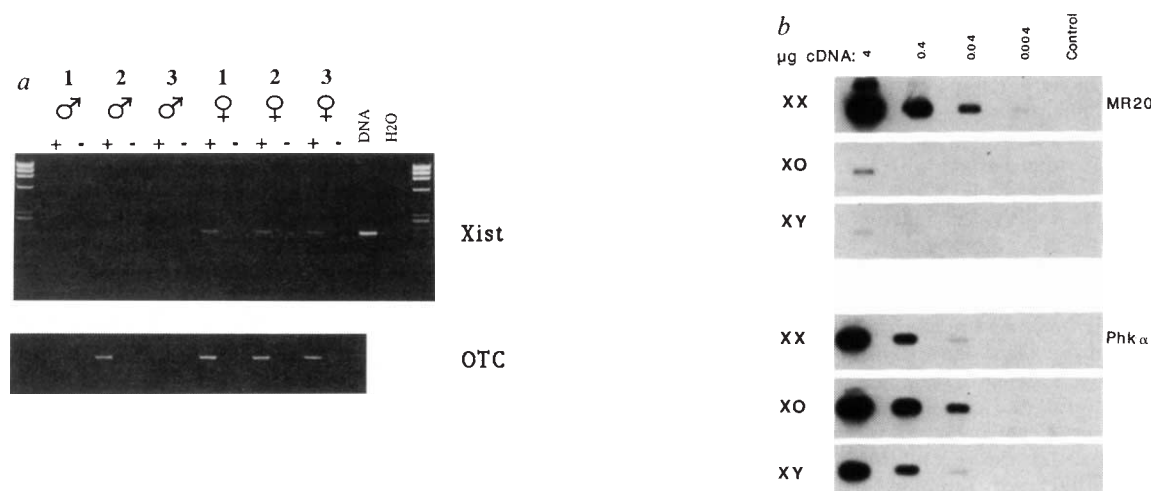


FIG. 2 Expression studies of the murine *Xist* gene. **a**, RT-PCR analysis of *Xist* expression in male and female mice. Reverse-transcribed liver RNA from six different mice (three males and three females) and mouse genomic DNA (lane 'DNA') were amplified using a set of primers from cDNA clone MR20 sequence. Plus and minus signs indicate the presence or absence respectively of reverse transcriptase in the reaction. *PhiX-HaeIII* digested DNA has been used for size calibration (outer lanes). An amplification product of the expected size (203 bp) can only be detected in female RT⁺ samples and in mouse genomic DNA with the PCR conditions used. A PCR with primers from the OTC cDNA sequence¹⁹ was performed as a positive control for the quality of the cDNAs and is shown at the bottom. Absence of amplification in the RT samples rules out possible contamination of the samples with genomic DNA. 'H₂O' lane shows PCR without template DNA as a test for PCR contamination. **b**, Quantitative analysis of *Xist* expression in XX, XO and XY samples. Serial dilutions of cDNA obtained from liver RNA of XX, XO and XY mice were amplified using a set of primers from the 5' end of the *Xist* gene. The PCR products (251 bp) were then transferred to a nylon membrane and hybridized with probe MR20. As an internal control we used a set of primers, amplifying a 240-bp fragment from the *Phka* gene, and a *Phka* cDNA probe²⁰.

METHODS. RT-PCR: total cellular RNA was prepared from liver of adult male and female mice by the guanidium thiocyanate method²¹ and 5 μg were

reverse-transcribed with murine Moloney leukaemia virus reverse transcriptase using random hexamer primers²². One-fiftieth of the cDNA was used for a PCR²³ amplification with 0.5 μmol primers for 40 cycles (94 °C for 1 min; 55 °C for 1 min; 72 °C for 1 min) with Cetus *Taq* polymerase and buffer. Primers used were: 594-1R (5' GGGACCTAACTGTTGGCTTTATCAG 3') and 594-1F (5' GAAGTGAATTGAAGTTTGGTCTAG 3'). Quantitative RT-PCR: 20 μg total liver RNA, isolated using the procedure described above, were treated with RQ1 DNase (Promega) before extraction with phenol and precipitation with ethanol. The RNA was then reverse-transcribed with 10 units AMV reverse transcriptase and 2 μg random hexamer primers²². Samples were treated with DNase-free RNase (10 μg) and precipitated with ethanol. For each RNA sample, a control experiment was carried out in parallel omitting the reverse transcriptase. Samples containing from 4 μg to 0.004 μg of cDNA were amplified for 20 cycles using Cetus *Taq* polymerase and 1 μmol of the primers (94 °C for 1 min; 54 °C for 1 min; 72 °C for 1 min) in a total volume of 50 μl. *Xist* primers used: 588-2F (5' ATCTAAGACAA-AATACATCATTCG 3') and 588-2R (5' CTTGGACTTAGCTCAGGTTTGTGTC 3'). *Phka* primers²⁰: A (5' AATTCACACTGCCAGGGCTTCAAC 3') and B (5' GCTTCAGCTCAGCTGGGTATAGTAT 3'). Thirty μl of each reaction product were electrophoresed on a 2% agarose gel, transferred to Hybond N+ (Amersham) and hybridized to random-primed MR20 *Xist* cDNA probe. The exposure time was 4 days, without intensifying screens.

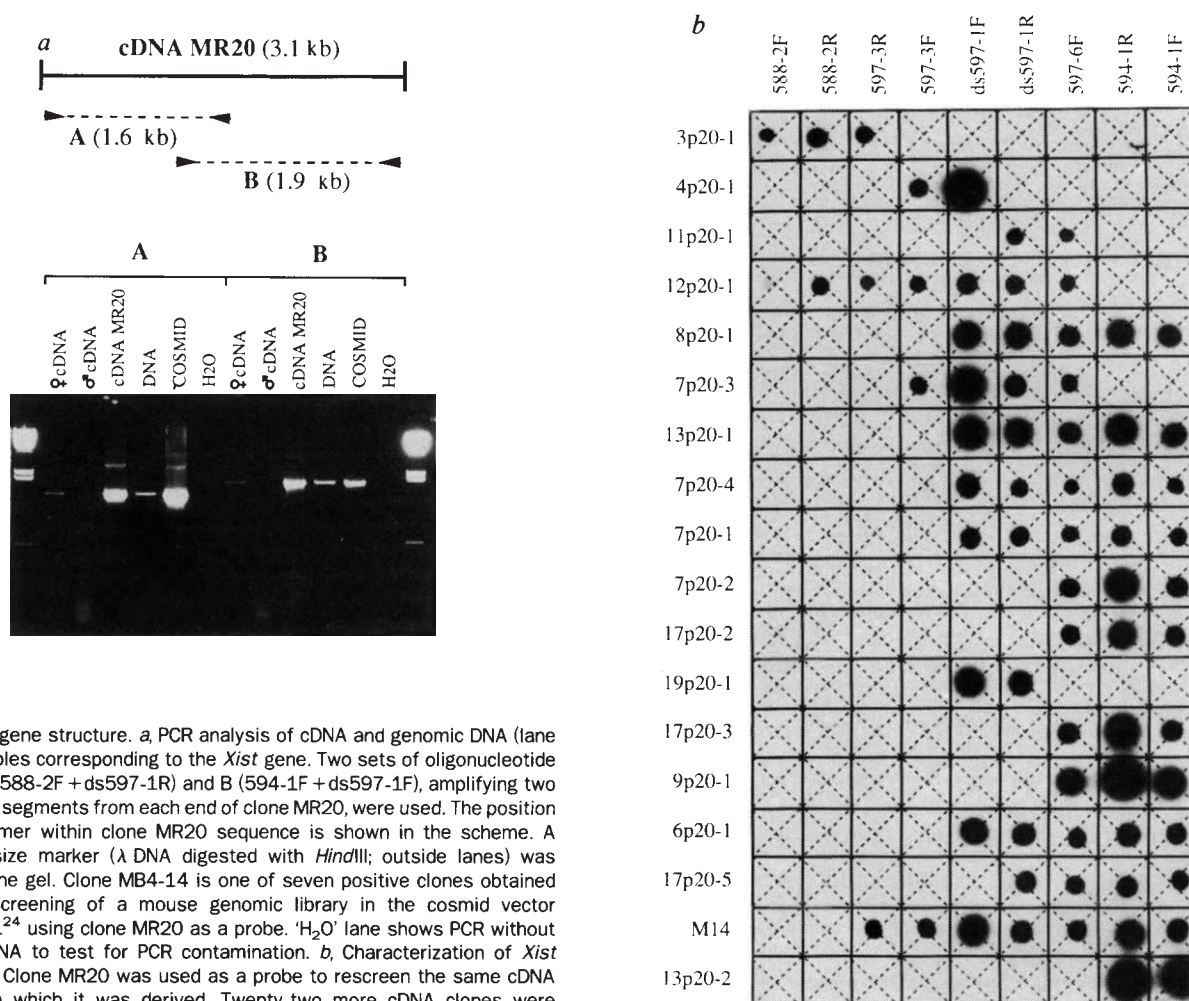


FIG. 3 *Xist* gene structure. *a*, PCR analysis of cDNA and genomic DNA (lane 'DNA') samples corresponding to the *Xist* gene. Two sets of oligonucleotide primers, A (588-2F + ds597-1R) and B (594-1F + ds597-1F), amplifying two overlapping segments from each end of clone MR20, were used. The position of each primer within clone MR20 sequence is shown in the scheme. A molecular size marker (λ DNA digested with *Hind*III; outside lanes) was loaded on the gel. Clone MB4-14 is one of seven positive clones obtained from the screening of a mouse genomic library in the cosmid vector pCOS2EMBL²⁴ using clone MR20 as a probe. 'H₂O' lane shows PCR without template DNA to test for PCR contamination. *b*, Characterization of *Xist* transcripts. Clone MR20 was used as a probe to rescreen the same cDNA library from which it was derived. Twenty-two more cDNA clones were identified and plaque-purified. Insert DNA from 18 cDNA clones was obtained by T7-T3 PCR amplification and spotted on a filter with a numbered grid. The name of each cDNA clone is indicated on the left. Vertical strips of the filter were cut and each strip was hybridized to an oligonucleotide located in a different position in the clone MR20 sequence. The nine oligonucleotide probes are indicated at the top in 5'-3' order.

METHODS. 6×10^5 plaque-forming units of a λ ZAPIII cDNA library were screened using the human *XIST* cDNA clone 14A (previously isolated by the immunoscreening of a λ GT11 cDNA library using anti-STS antibodies²⁵). Hybridization conditions were 65 °C in 1% SDS, 1M NaCl and 5% dextran sulphate. Washings were for 3 $\times$ 30 min at 65 °C in 3 $\times$ SSC, 0.05% SDS. The *Xist* cDNA clone MR20 was used as probe to rescreen the same library using identical conditions, except that the washings were at 65 °C in 0.2 $\times$ SSC. The latter conditions were also used for the screening of a mouse cosmid library (6×10^5 colony-forming units) in vector pCOS2EMBL with probe MR20. Two sets of oligonucleotide primers amplifying two overlapping segments from each end of clone MR20, were used for PCR analysis. Primers were 588-2F (5' ATCTAAGACAAAATACATCATTCCG 3') + ds597-1R (5' CACTGACTTGAAGTTACAGTAGGC 3') and 594-1F (5' GAAGTGAATTGAAGTTTGGTCTAG 3') + ds597-1F (5' GGTTTCATGGTCTTGTTCATATTCC 3'). PCR amplification was with 0.5 μ mol primers for 40 cycles (94 °C for 1 min; 54 °C for 1 min; 72 °C for 6 min) with Cetus *Taq* polymerase and buffer. Oligonucleotide hybridization and washings: insert DNA from the cDNA clones was amplified using T3 and T7 primers (36 cycles of 94 °C for 1 min; 50 °C for 1 min; 72 °C for 6 min). 3 μ l of a 1 in 10 dilution of the PCR products from each clone were spotted on a Genescreen Plus (NEN-Dupont) membrane. The membrane was then denatured in 0.4 M NaOH, neutralized in 0.2 M Tris, pH 7.5, 2 $\times$ SSC and baked at 80 °C. The nine oligonucleotides used in the analysis were: 588-2F, 588-2R, 597-3R (5' GTAATTTTCA-TTGTTCGATTGCCC 3'), 597-3F (5' CTGACATTGTTTCCCCTAACAAC 3'), ds597-1F, ds597-1R, 597-6F (5' TTAACGTACTCTCCCATATTGGG 3'), 594-1R and 594-1F (see above). 100 ng of each oligonucleotide were labelled using 10 μ Ci of [γ -³²P]ATP and T4 polynucleotide kinase (Pharmacia). Hybridization was for 8 h at 43 °C in 1% SDS, 1M NaCl, 5% dextran sulphate; washing was at 42 °C in 2 $\times$ SSC, 0.1% SDS. Filters were exposed overnight with intensifying screens.

1,500–2,000 of the mouse sequence. This region contains several subregions with identity between mouse and human sequence of >80%. PCR primers designed from one of these subregions gave PCR products from feline and bovine DNA which were then subcloned and sequenced. Figure 4b shows the alignment between mouse, man, feline and bovine sequences. Homology with other species, including rabbit, chimpanzee, wallaby, chicken and carp, was also found by low-stringency hybridization using both the human and the mouse probes (data not shown).

There is evidence that genes which escape X inactivation in man, such as *STS*¹³, *ZFX*¹⁴ and *RPSX4*¹⁵, are inactivated in the mouse (ref. 16, and A. Ashworth, personal communication), suggesting that there are important differences in X inactivation between man and mouse. *Xist* is the only gene known to be expressed from the murine inactive X chromosome and the only one among those also expressed from the human inactive X chromosome to show an identical expression pattern in the two species. *Xist* therefore shows novel features compared with other murine genes studied so far. Like its human homologue⁶, the murine *Xist* gene localizes to the critical mapping interval to which the *Xce* locus required (in *cis*) for X inactivation has been assigned, and its expression is female-specific in chromosomally normal individuals. The conservation of these features between two species as distantly related as man and mouse, suggests a functional role for this gene in female mammals generally. Besides its putative role in X inactivation, the isolation of the *XIST* gene, first in man and now in the mouse, offers an ideal molecular tool for the study of X inactivation and a starting point for the characterization of the X inactivation centre region in the two species. □



FIG. 4 *Xist* sequence analysis and comparison with other species. **a**, Nucleotide sequence of *Xist* cDNA. The sequence has been derived mainly from cDNA clone MR20 (bases 975–4,057) and partly from a *HindIII* subclone of cosmid MB4-14 (1–974). RT-PCR of female RNA confirmed that the portion of the sequence deriving from genomic DNA is also transcribed (data not shown). The predicted amino-acid sequence from the 894-bp open reading frame located at the 5' end of the sequence is also shown. Arrows indicate different units of the 24-bp repeat. **b**, Comparison between mouse, human, feline and bovine sequences. Oligonucleotide primers were designed from a region of high homology between the human and mouse sequences. These primers amplify a product in both human (257 bp) and mouse DNA (250 bp) and were used to amplify genomic DNA from cat and cow. A PCR product of 200 bp and of 203 bp was identified in cat and cow, respectively, and was subcloned and sequenced. The nucleotide sequences of this region in the four species are compared. Upper-case letters, aligned non-identical bases; lower-case letters, unaligned bases; dashed lines, aligned identical bases; dotted lines, gap.

METHODS. *λ* cDNA clones were subcloned in plasmids using the λ ZAPII plasmid rescue procedure according to the manufacturer's specifications (Stratagene). All nucleotide sequences were determined automatically using an Applied Biosystem ABI 370A fluorescent sequencer, as described by Gibbs *et al.*²⁶, and manually using a Sequenase sequencing kit (USB). Assembly of DNA sequences and search for open reading frames used the computer program DNA Strider²⁷. The sequence databases were searched using a previously described program²⁸. Multiple sequence alignment was according to ref. 29. Primers used for PCR were derived from the region of high homology and were TOP1 (5' caggatcctTCACATCTTCTCCACTTGAGA 3') and TOP2 (5' caggatcctTGTCTAATTTCTTCTCATTTGG 3'), the lower-case letters indicating a *Bam*HI cloning tail. Genomic DNA (500 μ g) from each species was amplified. PCR conditions were eight cycles at 94 °C for 1 min; at 50 °C for 1 min; at 72 °C for 1 min, followed by 27 cycles at 94 °C for 1 min; 53 °C for 1 min; 72 °C for 1 min. PCR products were digested with *Bam*HI followed by phenol-chloroform extraction and precipitation with ethanol. The fragments were subcloned in M13mp18 vector predigested with *Bam*HI, and subclones sequenced as described above.

Received 21 February; accepted 9 April 1991.

1. Brown, C. J. *et al.* *Nature* **349**, 38–44 (1991).
2. Gartler, S. M. & Riggs, A. D. *Rev. Genet.* **17**, 155–190 (1983).
3. Grant, S. G. & Chapman, V. M. *Rev. Genet.* **22**, 199–233 (1988).
4. Lyon, M. F. *Nature, new Biol.* **232**, 229–232 (1971).
5. Cattanach, B. M. *Rev. Genet.* **9**, 1–18 (1975).
6. Brown, C. J. *et al.* *Nature* **349**, 82–84 (1991).
7. Mitchell, M. D. *et al.* *Genetics* **121**, 803–809 (1989).
8. Keer, J. T. *et al.* *Genomics* **7**, 566–572 (1990).
9. Cattanach, B. M. *et al.* *Cytogenet. Cell Genet.* (in the press).
10. Amar, L. C. *et al.* *Genomics* **2**, 220–230 (1988).
11. Avner, P. R. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **84**, 5330–5334 (1987).
12. Sefton, L., Arnaud, D., Goodfellow, P., Simmler, M. C. & Avner, P. *Mammalian Genome* (in the press).
13. Mohandas, T. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **84**, 4954–4958 (1987).
14. Schneider-Gadick, A., Beer-Romero, P., Brown, L. G., Nussbaum, R. & Page, D. C. *Cell* **57**, 1247–1258 (1989).
15. Fisher, E. M. C. *et al.* *Cell* **63**, 1205–1218 (1990).
16. Jones, J., Peters, J., Rasberry, C. & Cattanach, B. M. *Genet. Res.* **53**, 193–199 (1989).
17. Church, G. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **81**, 1991–1995 (1984).
18. Feinberg, A. P. & Volgestein, B. *Analyt. Biochem.* **132**, 6–13 (1983).
19. Scherer, S. E., Veres, G. & Caskey, C. T. *Nucleic Acids Res.* **16**, 1593–1601 (1988).
20. Bender, P. K. & Lalley, P. A. *Proc. natn. Acad. Sci. U.S.A.* **86**, 9996–10000 (1989).
21. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. *Biochemistry* **18**, 5294–5299 (1979).
22. Kawasaki, E. S. in *PCR Protocols. A Guide to Methods and Applications* (eds Innis, A., Gelfond, D. H., Sninsky, J. J. & White, T. J.) (Academic, New York, 1990).
23. Saiki, R. K. *et al.* *Science* **239**, 487–491 (1988).
24. Ehrlich, E., Craig, A., Poustka, A., Frischauf, A. M. & Lehrach, H. *Gene* **57**, 229–237 (1987).
25. Ballabio, A. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **84**, 4519–4523 (1987).

26. Gibbs, R. A., Nguyen, P. N., McBride, L. J., Koepf, S. M. & Caskey, C. T. *Proc. natn. Acad. Sci. U.S.A.* **86**, 1919–1923 (1989).
 27. Marck, C. *Nucleic Acids Res.* **16**, 1829–1836 (1988).
 28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. molec. Biol.* **215**, 403–410 (1990).
 29. Feng, D. F. & Doolittle, R. F. *J. molec. Evol.* **25**, 351–360 (1987).

ACKNOWLEDGEMENTS. We thank D. Toniolo for the mouse cosmid library, C. Brown and J. Rupert for unpublished sequence data and D. Ledbetter and C. Chinnault for critical reading of the manuscript. Sequence comparison was performed using the facilities of the Molecular Biology Information Resource, Baylor College of Medicine. The nucleotide sequence data reported will appear in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under the accession number X59289 MURINE X1ST cDNA.

Conservation of position and exclusive expression of mouse *Xist* from the inactive X chromosome

Neil Brockdorff, Alan Ashworth*, Graham F. Kay, Penny Cooper, Sandy Smith, Veronica M. McCabe, Dominic P. Norris, Graeme D. Penny, Dipika Patel & Sohaila Rastan†

Section of Comparative Biology, MRC Clinical Research Centre, Harrow HA1 3UJ, UK

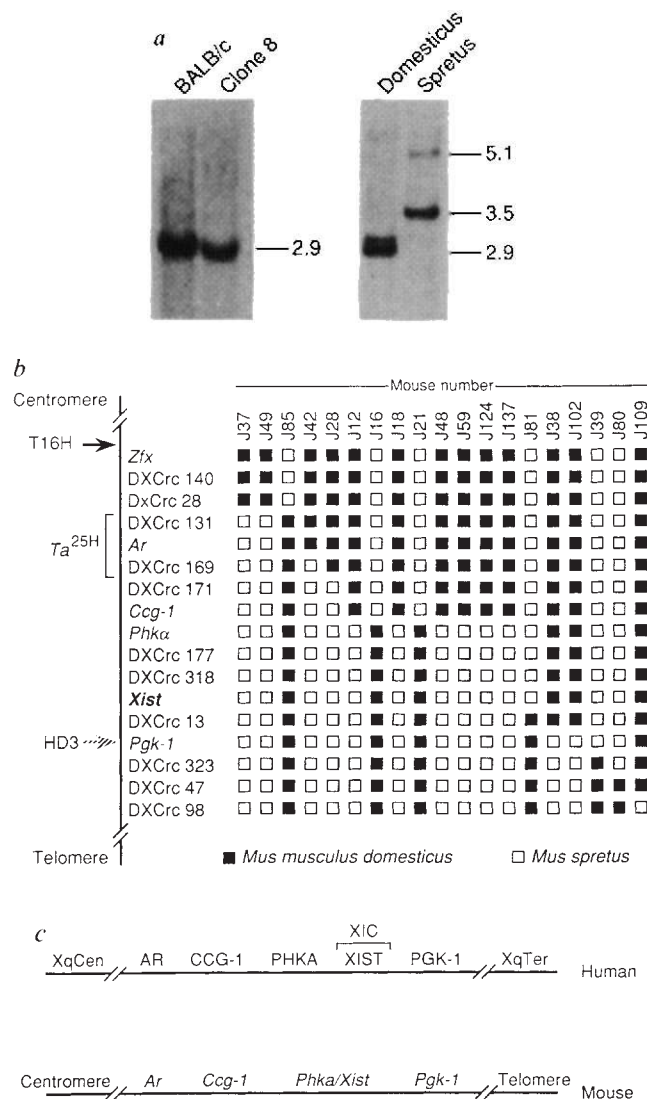
* Chester Beatty Laboratories, The Institute of Cancer Research, London SW3 6JB, UK

† To whom correspondence should be addressed

X-CHROMOSOME inactivation in mammals is a regulatory phenomenon whereby one of the two X chromosomes in female cells is genetically inactivated, resulting in dosage compensation for X-linked genes between males and females¹. In both man and mouse, X-chromosome inactivation is thought to proceed from a single *cis*-acting switch region or inactivation centre (XIC/Xic)^{2–5}. In the human, XIC has been mapped to band Xq13 (ref. 6) and in the mouse to band XD (ref. 7), and comparative mapping has shown that the XIC regions in the two species are syntenic⁸. The recently described human *XIST* gene maps to the XIC region⁶ and seems to be expressed only from the inactive X chromosome⁹. We report here that the mouse *Xist* gene maps to the Xic region of the mouse X chromosome and, using an interspecific *Mus spretus*/*Mus musculus domesticus* F₁ hybrid mouse carrying the T(X; 16)16H translocation, show that *Xist* is exclusively expressed from the inactive X chromosome. Conservation between man and mouse of chromosomal position and unique expression exclusively from the inactive X chromosome lends support to the hypothesis that *XIST* and its mouse homologue are involved in X-chromosome inactivation.

We have used a 1.3-kilobase (kb) human probe, generated by the polymerase chain reaction from the published human *XIST* sequence, to screen an oligo(dT)-primed complementary

FIG. 1 Genetic mapping of the mouse *Xist* gene. *a*, Hybridization of the 2.7-kb insert from the *Xist* cDNA to a Southern blot of *TaqI*-restricted DNA from a BALB/c mouse and from the clone 8 hybrid cell line¹⁷ carrying only the X chromosome of mouse together with human HeLa cell chromosomes. The *Xist* gene maps to the mouse X chromosome (panel 1). A single band corresponding to 2.9 kb was detected in both BALB/c and clone 8 DNA. For interspecific backcross pedigree analysis, a *TaqI* restriction fragment-length variant between *Mus musculus domesticus* (2.9 kb) and *Mus spretus* (5.1 and 3.5 kb) was detected (panel 2). An additional 3.1-kb weak band, cosegregating with the 2.9-kb band, was seen in the *domesticus* mice used for the backcross. This band was not seen in BALB/c or clone 8 DNA, and is thought to represent a polymorphism in the genetic background (strains 101 and C3H) of the *domesticus* parent used for the backcross. Molecular sizes are indicated in kbp. *b*, The *TaqI* restriction-fragment length variant between *domesticus* and *spretus* was used to map *Xist* to the Xic region of the mouse X chromosome. This region is delineated by the breakpoint



in T16H (proximal limit) and the deletion breakpoint in embryonic stem cell line HD3 (the distal limit)⁷. A small region within these limits, deleted in the *Ta^{25H}* mutation, is excluded as a candidate region for the Xic (refs 18, 19). A panel of 17 interspecific backcross mice with recombinant break points within Xic were used to map the *Xist* gene with respect to several other molecular markers in this region¹⁰. This analysis locates the *Xist* gene between *Ccg-1* and the *DXCrc13* loci, and shows that it cosegregates with *Phka*, *DXCrc177* and *DXCrc318*. The haplotypes of the recombinant X chromosome of the 17 backcross progeny are shown for each of the probes used. *c*, A comparative map of the X-inactivation centre region of the human and mouse X chromosome illustrates that the genetic map position of *Xist* with respect to flanking markers is identical in the two species.

METHODS. The human *XIST* probe used to screen the mouse cDNA library was a 1.3-kb fragment generated from the published human *XIST* sequence⁹ by PCR from HL60 cell line cDNA. The primers used were AAG-GTGAAGGCTCATAGG and CTGCATGATTGCCAATACAC, corresponding to nucleotides 121–140 and 1,462–1,443 of the human sequence⁹. The cDNA library, an oligo (dT)-primed library from 17.5-day-old mouse embryos (Clontech), was screened at low stringency (5×SSC, 10% dextran sulphate, 1% SDS and 100 μg ml⁻¹ salmon sperm DNA, at 50 °C overnight; wash conditions: 2×SSC for 2×15 min at room temperature followed by 2×SSC, 1% SDS for 2×30 min at 50 °C). A single positive hybridizing clone with a 2.7-kb insert was obtained. Partial sequence analysis showed that it coded for the murine homologue of *XIST*. This clone overlaps with part of the published human sequence and shows about 75% sequence homology; multiple termination codons were present in all reading frames (data not shown). Southern hybridizations were carried out under standard conditions¹⁷. The production of the interspecific backcross progeny and the detailed molecular mapping of the 17 backcross progeny used in the mapping panel is described elsewhere (ref. 10, and G.F.K., R. V. Thakker and S.R., manuscript submitted).

DNA library from 17.5-day-old mouse embryos. A single positive clone containing a 2.7-kb insert was isolated, which corresponded to part of the mouse *Xist* gene. We demonstrated that the *Xist* cDNA maps to the mouse X chromosome by analysing a human-mouse somatic cell hybrid, clone 8, which carried only the X chromosome of mouse, together with HeLa-cell chromosomes (Fig. 1a, panel 1). To define the position of *Xist* on the mouse X chromosome, *Mus musculus domesticus*/*Mus spretus* interspecific backcross pedigree analysis was used. A *TaqI* restriction fragment-length variant between *domesticus* and *spretus* was found for the *Xist* probe (Fig. 1a, panel 2). The segregation of this variant was then analysed through a panel of 17 interspecific backcross progeny. These progeny had been previously characterized with many molecular and genetic markers in the Xic region of the mouse X chromosome¹⁰. The progeny were selected on the basis of their having recombination breakpoints between *Zfx* (the proximal limit of the Xic region) and *DXCrc98* (the distal limit of the Xic region)¹⁰ (Fig. 1b). The *Xist* gene maps in the position shown in Fig. 1b between *Ccg-1* and *DXCrc13*, and cosegregates with *Phk α*, *DXCrc177* and *DXCrc318*. Comparison of maps of the XIC/Xic region on the human and mouse X chromosome shows that the position of *Xist* with respect to flanking markers is identical in the two species (Fig. 1c).

We examined the expression of the *Xist* gene in the mouse by northern blotting (Fig. 2). *Xist* messenger RNA was detected in females only, suggesting that expression is confined to the inactive X chromosome, as seems to be the case for the human *XIST* gene⁹. The strong heterogeneous signal in female RNA

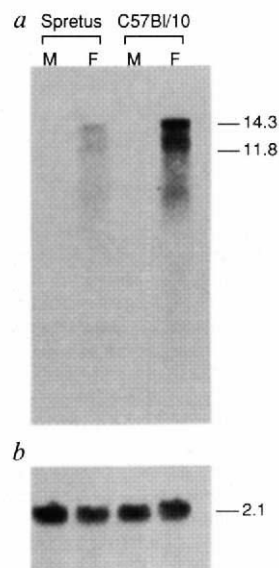


FIG. 2 Expression of the *Xist* gene in male and female mice. *a*, Northern blot of total cellular RNA from kidney of male (M) and female (F) C57Bl/10 (*Mus musculus domesticus*) and *Mus spretus* mice hybridized with the 2.7-kb mouse *Xist* cDNA. The probe hybridizes only to the female RNA samples, suggesting that the transcript is produced from the inactive X chromosome. Major bands are detected at positions corresponding to 14.3 and 11.8 kb. *b*, The same northern blot filter stripped and reprobed with mouse actin cDNA, demonstrating that RNA is present and intact in all lanes. Molecular sizes are indicated in kb.

METHODS. Total cellular RNA was prepared using the guanidium thiocyanate method²⁰. Northern blots were carried out using the glyoxal method²¹. Total RNA (20 µg) was electrophoresed on a 0.8% agarose gel and then transferred to nylon membranes. The blots were hybridized with ³²P-labelled *Xist* cDNA insert in 50% formamide, 5×SSC, 10% dextran sulphate, 1× Denhardt's solution, 1% SDS and 100 µg ml⁻¹ denatured salmon sperm DNA at 42°C overnight. Wash conditions, 2×SSC for 2×15 min at room temperature followed by 2×SSC, 1% SDS for 2×20 min at 65°C. The actin probe was a 1.35-kb mouse cDNA²².

indicates a series of many different transcripts, showing size and/or conformation heterogeneity, as for the human gene. But large discrete bands at positions corresponding to 14.3 kb and 11.8 kb could be clearly seen (Fig. 2a), indicating that there are two predominant transcription products of *Xist*. *Xist* expression was examined in both *Mus musculus domesticus* (C57Bl/10 strain) mice and *Mus spretus* mice (Fig. 2a). The level of expression was markedly lower in the *Mus spretus* female than in the C57Bl/10 female (Fig. 2a), and reprobing with actin cDNA (Fig. 2b) confirmed that equal quantities of intact RNA had been loaded in each lane. The *Xce* locus, which maps to the Xic region^{11,12}, has been proposed as a candidate for Xic in the mouse because different alleles affect the randomness of X-inactivation^{13,14}. That *Xist* expression in *Mus spretus* is less than that in C57Bl/10 mice is intriguing, because *Mus spretus* seems to have a strong *Xce* allele; that is, the *Mus spretus* X chromosome is more likely to remain active in F₁ interspecific hybrids between *Mus spretus* (spe) and *Mus musculus domesticus* (dom)¹⁵.

To prove that expression of *Xist* was from the inactive X chromosome, rather than from the active X chromosome in response to the presence of the inactive X chromosome, we examined *Xist* expression using an F₁ interspecific hybrid mouse between *Mus spretus* (spe) and *Mus musculus domesticus* (dom) carrying the T(X;16)16H translocation (T16) (ref. 15). These animals provide an *in vivo* system to assess whether *Xist* is expressed from the active (T16^{dom}) or inactive (X^{spe}) chromosome. This translocation causes nonrandom inactivation of the X^{spe} chromosome in T16^{dom}/X^{spe} mice¹⁶: only the T16^{dom} alleles of genes known to be X-inactivated are expressed¹⁵. These animals provide an *in vivo* system to assess whether *Xist* is expressed from the active (T16^{dom}) or inactive (X^{spe}) chromosome.

Control interspecific X^{dom}/X^{spe} females, in which either X chromosome may be inactivated, express both *Xist*^{dom} and



FIG. 3 *Xist* is only expressed from the inactive X chromosome in mice. RNA prepared from the livers of X^{dom}/X^{dom}, X^{spe}/X^{spe}, X^{dom}/X^{spe} and T16^{dom}/X^{spe} mice was reverse-transcribed into cDNA. *Xist* cDNA was amplified using gene-specific primers and the PCR products directly sequenced. *Xist*^{spe} differs from *Xist*^{dom} by the presence of 2 bp at the position indicated. In the control interspecific F₁ animal (X^{dom}/X^{spe}) both alleles are expressed. Interspecific F₁ T16^{dom}/X^{spe} animals, however, express only the *Xist*^{spe} allele confirming that *Xist* is expressed exclusively from the inactive X chromosome.

METHODS. Generation of T16^{dom}/X^{spe} and X^{dom}/X^{spe} animals and analysis of gene expression were performed as described¹⁵. *Xist* cDNA was amplified with primers (TAAGGACTACTTAACGGGCT and TCACATCTGCTCCACTTGAG) designed on the basis of the partial sequence of the murine cDNA clone (not shown). The product of the PCR is 300 bp. Control PCR reactions on RNAs incubated without reverse transcriptase demonstrated that amplification was not due to contaminating genomic DNA. Sequence variants between *Mus musculus domesticus* and *Mus spretus* were identified by DNA sequencing. The sequence shown is of *Xist* mRNA and is the opposite strand to that shown on the sequencing gel.

Xist^{spe} transcripts (Fig. 3). The relative expression of *Xist*^{dom} is higher than *Xist*^{spe}. This is evident in the C track which allows a within-track comparison, where the CC^{dom} doublet is significantly stronger than the CC^{spe} doublet. This result supports our northern-blot analysis which showed an inverse correlation between the level of *Xist* expression and the strength of the *Xce* allele (Fig. 2). By contrast, T16^{dom}/X^{spe} mice, in which the X^{spe} is always the inactive X chromosome, express only the *Xist*^{spe} allele (Fig. 3). Thus, expression of *Xist* is exclusively from the inactive X chromosome *in vivo*, and is not from the active X chromosome in response to a signal from the inactive X chromosome(s).

The conservation of map position between human *XIST* and mouse *Xist*, together with their exclusive expression from the inactive X chromosome, strongly supports the hypothesis that they are involved in X-chromosome inactivation. □

Received 13 February; accepted 23 April 1991.

1. Lyon, M. F. *Nature* **190**, 372–373 (1961).
2. Russell, L. B. *Science* **140**, 976–978 (1963).
3. Cattanach, B. M. A. *Rev. Genet.* **9**, 1–18 (1975).
4. Mattei, M. G., Mattei, J. F., Vidal, I. & Giraud, F. *Hum. Genet.* **56**, 401–408 (1981).
5. Therman, G., Sarto, G. E. & Patau, K. *Chromosoma* **44**, 361–366 (1974).
6. Brown, C. J. *et al.* *Nature* **349**, 82–84 (1991).
7. Rastan, S. & Robertson, E. J. *J. Embryol. exp. Morph.* **90**, 379–388 (1985).
8. Keer, J. T. *et al.* *Genomics* **7**, 566–572 (1990).
9. Brown, C. J. *et al.* *Nature* **349**, 38–44 (1991).
10. Brockdorff, N. *et al.* *Genomics* **10**, 17–22 (1991).
11. Cattanach, B. M., Rasberry, C. & Andrews, S. J. *Mouse News Letter* **83**, 165 (1989).
12. Rastan, S. & Brown, S. D. M. *Genet. Res.* **56**, 99–106 (1990).
13. Cattanach, B. M. & Papworth, D. *Genet. Res.* **38**, 57–70 (1981).
14. Johnston, P. G. & Cattanach, B. M. *Genet. Res.* **37**, 151–160 (1981).
15. Ashworth, A., Rastan, S., Lovell-Badge, R. & Kay, G. *Nature* (in the press).
16. Lyon, M. F., Searle, A. G., Ford, C. E. & Ohno, S. *Cytogenetics* **3**, 306–323 (1964).
17. Brockdorff, N., Montague, M., Smith, S. & Rastan, S. *Genomics* **7**, 573–578 (1990).
18. Cattanach, B. M. *et al.* *Cytogenet. Cell Genet.* (in the press).
19. Brockdorff, N., Kay, G., Cattanach, B. M. & Rastan, S. *Mammalian Genome* (in the press).
20. Chirgwin, J. M. *et al.* *Biochemistry* **18**, 5294–5299 (1979).
21. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, New York 1989).
22. Minty, A. J. *et al.* *J. Biol. Chem.* **256**, 1008–1014 (1981).

ACKNOWLEDGEMENTS. We thank J. McVey for the actin cDNA probe, D. Shepherd and other animal husbandry staff at the MRC Clinical Research Centre, the medical illustration department at the centre, and S. Jenkins for typing the manuscript. This work was supported by the MRC and the Cancer Research Campaign.

Tetramerization of an RNA oligonucleotide containing a GGGG sequence

Joseph Kim, Chaejoon Cheong & Peter B. Moore

Department of Chemistry, Yale University, 225 Prospect Street, New Haven, Connecticut 06511, USA

POLY rG can form four-stranded helices¹. The Hoogsteen-paired quartets of G residues on which such structures depend are so stable that they will form in 5'-GMP solutions, provided that Na⁺ or K⁺ are present (see for example, refs 2–4). Telomeric DNA sequences, which are G-rich, adopt four-stranded antiparallel G-quartet conformations *in vitro*^{5,6}, and parallel tetramerization of G-rich sequences may be involved in meiosis^{7,8}. Here we show that RNAs containing short runs of Gs can also tetramerize. A 19-base oligonucleotide derived from the 5S RNA of *Escherichia coli* (strand III), 5'GCCGAUGGUAGUGGGGU3', forms a K⁺-stabilized tetrameric aggregate that depends on the G residues at its 3' end. This complex is so stable that it would be surprising if similar structures do not occur in nature.

Strand III, which is prepared from the 5S RNA of *E. coli* nucleolytically⁹, forms an aggregated species called III*, which migrates much more slowly than strand III on polyacrylamide gels. This III* species is very stable; unlike ordinary double

helices it persists in low-salt buffers containing 8 M urea. Recent experiments demonstrated that III* is preferentially stabilized by K⁺, suggesting that it might be a four-stranded structure, like the several G-rich DNA structures reported recently, which are also K⁺-stabilized^{5–8}.

If III* is a four-stranded G structure, the residues most likely to be responsible for aggregation are the four G residues near the 3' end of strand III, and strand III should be able to aggregate with any RNA that includes a run of four (or more) G residues, as has been demonstrated for some DNA tetramers⁸. An RNA hexamer was synthesized to test this prediction (5'UGGGGU3'), and equimolar mixtures of hexamer and strand III were analysed by PAGE.

If heterotetramers form in these mixtures, as expected, this experiment could have two different outcomes, depending on whether the aggregates formed are parallel stranded^{7,8,10} or antiparallel stranded^{5,6}. Parallel aggregation will generate five tetrameric species distinguishable on gels (Fig. 1), but antiparallel tetramerization should give rise to six. The 'extra' antiparallel species arises because the critical G-run in strand III is near one end of the molecule. Consequently there are two possible antiparallel III₂-H₂ tetramers, one with its strand III components parallel and the other with them antiparallel, and they should be distinguishable electrophoretically because they differ in hydrodynamic radius.

Only five species can be detected in mixtures of strand III and hexamer in addition to their single-stranded forms: H₄, III-H₃, III₂-H₂, III₃-H and III₄ (Fig. 2). We conclude therefore that III* is tetrameric, that the interactions stabilizing it involve residues at its 3' end, and that III* is parallel-stranded. Hexamer tetramerizes with itself, as expected.

If III* is stabilized primarily by interactions involving the four G residues at its 3' end, it should be possible to convert the 5' portions of each oligonucleotide in III* into double helix by hybridization with an RNA of appropriate sequence without

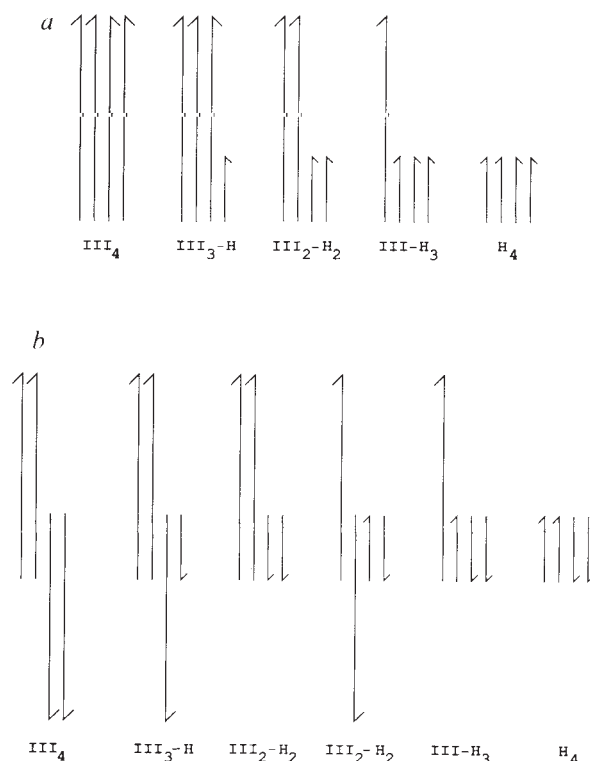


FIG. 1 The dependence of the number of tetrameric aggregates expected on relative strand orientation. The tetramers that can form in mixtures of hexamer and strand III depend on whether the structures in question are parallel stranded (a) or antiparallel (b), as shown.

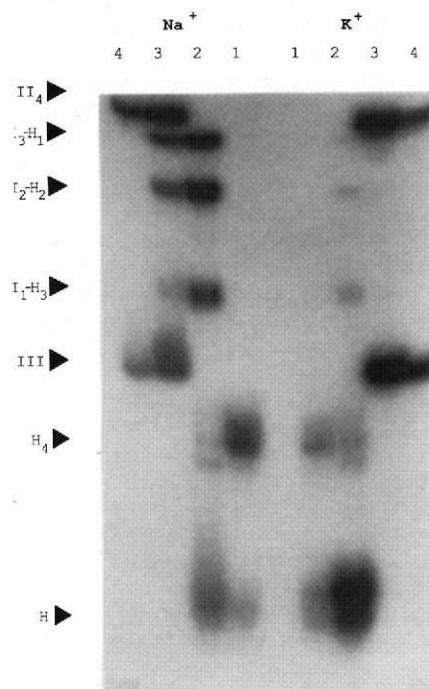


FIG. 2 The formation of heterotetramers from mixtures of strand III and hexamer. Autoradiogram of a nondenaturing, 20% polyacrylamide gel run in 50 mM Tris-borate, pH 8, 1 mM EDTA for ~8 h at 9 V cm^{-1} . All samples were dissolved in $10 \mu\text{l}$ of 1 mM EDTA, 10 mM Tris-HCl, pH 8.2, and 0.1 M NaCl (left side of gel) or 0.1 M KCl (right side). Lane 1, 5 nmol hexamer plus a small amount of hexamer ^{32}P -labelled at its 5' end; lane 2, 5 nmol hexamer plus 5 nmol strand III and ^{32}P -labelled hexamer; lane 3, as lane 2, except that the ^{32}P label is in strand III; lane 4, 5 nmol strand III plus ^{32}P -labelled strand III. (The amount of radioactive material included in lanes 1 and 4 was one quarter the amount included in lanes 2 and 3.) All eight samples, four in NaCl and four in KCl, were incubated at 95°C for 2 min, cooled quickly, and then incubated overnight at 4°C . The hexamer used was synthesized chemically, and deprotected as described by Webster and Spicer¹². Polynucleotides were end-labelled with ^{32}P -phosphate using polynucleotide kinase. It is evident from lanes 1 and 2 that the hexamer preparation used was impure. Purification of hexamer was unusually difficult because it aggregates so strongly. It is likely that the 'short' material is hexamer lacking its terminal U residue.

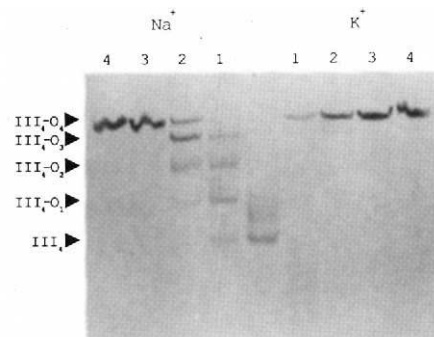


FIG. 3 The formation of complexes between III* and octamer. The strand III concentration was $120 \mu\text{M}$ in all samples. Lanes 1-4 contain 40, 80, 160 and $240 \mu\text{M}$ octamer (5'-CCAUCGCC-3'), respectively, giving molar ratios of strand III to octamer of ~4:1, 2:1, 1:1 and 0.5:1. The middle lane contains III* only. Samples were made up in 50 mM Tris-HCl, pH 7.8, and either 100 mM NaCl or 100 mM KCl, heated at 40°C for 1 h, and then dried under vacuum before loading onto a 18% polyacrylamide gel in 50 mM Tris-borate, pH 8.2, 1 mM EDTA. Gels were run at 25 V cm^{-1} for 21 h at 5°C , and bands were made visible by methylene-blue staining.

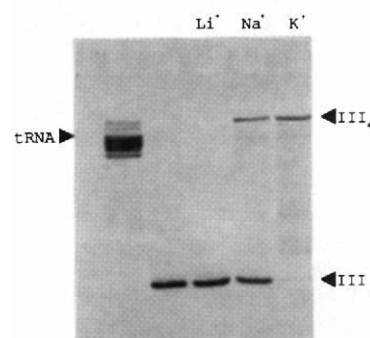


FIG. 4 The dependence of III* stability on monovalent cation identity. Preformed III* was incubated at 37°C for 90 min in 50 mM Tris-HCl, pH 7.8, either without salt or with 100 mM LiCl, 100 mM NaCl or 100 mM KCl at a concentration of $130 \mu\text{M}$. Samples were cooled to room temperature, mixed with an equal volume of formamide and heated to 60°C for 5 min before being loaded on a 16% polyacrylamide gel containing 8 M urea, 90 mM Tris-borate, pH 8.2, 0.5 mM EDTA. The gel was run at 20 V cm^{-1} for 4 h at room temperature, and bands were visualized by staining with methylene blue. III, strand III monomer; IIII, III*; tRNA, transfer RNA (76 bases).

causing it to disaggregate⁷. This was tested using an RNA octamer complementary to the 5' eight bases of strand III, 5'-CCAUCGCC3'. Octamer was added to III* at different ratios, and the molecular masses of the species in the mixtures that resulted were examined on gels (Fig. 3). As expected, four new species were generated, the complexes of III* with one, two, three or four molecules of octamer.

G-driven DNA tetramerization has a marked dependence on monovalent cations^{6,8}. K^+ stabilizes parallel and antiparallel DNA tetramers, but its effects on tetramerization depend on the pathway for tetramer formation. Antiparallel G-quartet DNA⁶, which is a unimolecular structure, forms in K^+ as readily as in Na^+ . But the formation of parallel G4-DNA, which is tetramolecular, is kinetically hindered by K^+ because it stabilizes unproductive, antiparallel intermediates⁸.

Preformed III* incubated in K^+ survives largely intact in gels containing 8 M urea, but is less stable on such gels if incubated in Na^+ , and dissociates completely if incubated in Li^+ (Fig. 4). On the other hand, III* forms faster in Na^+ than in K^+ at 37°C (at $100 \mu\text{M}$ strand III), and does not form at all in Li^+ under these conditions (data not shown). As the kinetics of tetramerization in K^+ are very slow (half-life $t_{1/2} > 48 \text{ h}$) even in the absence of a detectable intermediate, it is unclear whether the stability of III* displayed in Fig. 4 is a thermodynamic effect or a kinetic one.

The complexes formed by runs of Gs only four long are so stable that we would expect them to have been encountered many times before, but we know of only one analysis of such aggregates¹¹. □

Received 11 March; accepted 15 April 1991.

1. Zimmerman, S. B., Cohen, G. H. & Davies, D. R. *J. molec. Biol.* **92**, 181-192 (1975).
2. Zimmerman, S. B. *J. molec. Biol.* **106**, 663-672 (1976).
3. Pinnavaia, T. J. *et al.* *J. Am. chem. Soc.* **100**, 3625-3627 (1978).
4. Borzo, M., Detellier, C., Laszlo, P. & Paris, A. *J. Am. chem. Soc.* **102**, 1124-1134 (1980).
5. Sundquist, W. I. & Klug, A. *Nature* **342**, 825-829 (1989).
6. Williamson, J. R., Raghuraman, M. K. & Cech, T. R. *Cell* **59**, 871-880 (1989).
7. Sen, D. & Gilbert, W. *Nature* **334**, 364-366 (1988).
8. Sen, D. & Gilbert, W. *Nature* **344**, 410-414 (1990).
9. Moore, P. B. *et al.* *Meth. Enzym.* **164**, 158-174 (1988).
10. Jin, R., Breslauer, K. J., Jones, R. A. & Gaffney, B. L. *Science* **250**, 543-546 (1990).
11. Studencki, A. B. thesis, Univ. Illinois (1981).
12. Webster, K. & Spicer, E. K. *J. biol. Chem.* **265**, 19007-19014 (1990).

ACKNOWLEDGEMENTS. We thank the Yale University School of Medicine Protein and Nucleic Acid Chemistry Facility for synthetic RNA oligonucleotides. This work was supported by the NIH and a Damon Runyon-Walter Winchell Cancer Research Fund Fellowship (C.C.).